Convergence Rate Analysis of Non-I.I.D. SplitFed Learning with Partial Worker Participation and Auxiliary Networks

Amirreza Talebi *

Posted Date: 4 September 2024

doi: 10.20944/preprints202409.0335.v1

Keywords: SplitFed Learning; Convergence Theory; Federated Learning; Auxiliary Networks; Machine Learning

Article

Convergence Rate Analysis of Non-i.i.d. SplitFed Learning with Partial Worker Participation and Auxiliary Networks

Amirreza Talebi

The Ohio State University, Columbus, OH, USA; talebi.14@osu.edu

Abstract: In conventional Federated Learning (FL), clients work together to train a model managed by a central server, intending to speed up the learning process. However, this approach imposes significant computational and communication burdens on clients, particularly with complex models. Additionally, while FL strives to protect client privacy, the server's access to local and global models raises security concerns. To address these challenges, Split Learning (SL) separates the model into parts handled by the client and the server, though it suffers from inefficiencies due to sequential client participation. To overcome these issues, SplitFed Learning (SFL) was proposed, which combines the parallelism of FL with the model-splitting strategy of SL, enabling simultaneous training by multiple clients. Our main contribution is the theoretical analysis of SFL, which, for the first time, includes non-i.i.d. datasets, non-convex loss functions, and both full and partial client participation. We provide convergence proofs for a state-of-the-art SFL algorithm based on conventional convergence analysis assumptions for FL. Our results prove that we can recover the linear convergence rate of conventional FL for the SFL algorithm with the distinction that increasing the number of local steps or clients may not speed up the convergence in SFL.

Keywords: SplitFed Learning; Convergence Theory; Federated Learning; Auxiliary Networks; Machine Learning

1. Introduction

In the conventional Federated Learning (FL), several clients in parallel, train a model jointly particularly leading to speed-up in the learning process under the supervision of a server [1]. Hence, given a central server and N clients as participants in the training, an optimization problem of the below form is solved by FL:

$$\min_{\tilde{\mathbf{x}}\in\mathbb{R}^{\tilde{d}}} f(\tilde{\mathbf{x}}) \triangleq \min_{\tilde{\mathbf{x}}\in\mathbb{R}^{\tilde{d}}} \frac{1}{m}\sum_{i=0}^{m-1} F_i(\tilde{\mathbf{x}}) \tag{1}$$

In which, $F_i(\tilde{\mathbf{x}}) \triangleq \mathbb{E}_{\xi\sim D_i}[F_i(\tilde{\mathbf{x}},\xi)]$ can be a non-convex loss function and ξ corresponds to a random sample of local dataset of the client i, D_i. In the FL training, there are m clients training on their local datasets. However, FL encounters the challenge that clients must train the entire model, placing a considerable computational burden, particularly with complex and large-scale models. Additionally, gathering all client data and broadcasting the aggregated model at each round can result in substantial communication overhead. While one of the principal aims of FL is to safeguard clients' privacy, the server retains access to both the client's local and global models, prompting security concerns [2]. To address the computational limitations and further safeguard the privacy of the client-side model, [3] pioneered SL, dividing the ML model into two parts. The client trains one portion of the model, while the server trains the remaining portion. However, according to [2], this method incurs notable training time overhead, as only one client can engage in split learning (SL) at any given time, leaving others idle. To address this issue, they proposed SplitFed Learning (SFL), which integrates both the parallel computational capabilities of clients from FL and the benefits of split models from SL. In particular, the convergence theory of the SFL framework has not been thoroughly explored in the existing literature.

Our primary contribution lies in establishing the theoretical underpinnings for SFL, incorporating the general assumptions of traditional FL, non-convex loss functions, non-identically independently distributed (non-i.i.d.) datasets, and addressing both full and partial client participation in the SFL training process. Our proof is for the state-of-the-art algorithm for SFL developed by [4] based on conventional assumptions in FL settings. We demonstrate that the SFL can still recover the linear convergence rate of conventional FL. However, changes in the number of clients and local steps cannot speed up the convergence.

2. Related work

1. **SL and FL**

   The reference [5] introduces a personalized SL framework to address issues like data leakage and non-iid datasets in decentralized learning. It proposes an optimal cut layer selection method using multiplayer bargaining and the Kalai-Smorodinsky bargaining solution (KSBS). This approach efficiently balances the time of training, usage of energy, and privacy of data. Each device tailors its model for non-i.i.d. datasets while they have a common server-side model which ensures robustness by generalization. Simulation results validate the framework's effectiveness in achieving optimal utility and addressing decentralized learning challenges. However, they do not address the communication overhead caused by transmitting the forward-propagation results at each local step. The reference [6] provides convergence analysis for Sequential Split Learning (SSL), a variant of SL in which the model training process is conducted sequentially, with each client trained one after the other, on heterogeneous data. It compares SSL with Federated Averaging (FedAvg) showing SSL's superiority on extremely heterogeneous data. However, in practice, if the heterogeneity of data is mild, FedAvg outperforms SSL. Also, SSL still suffers from large communication overheads between the server and clients.
2. **SplitFed learning**

   The reference [7] presents AdaSFL, a method designed to optimize model training efficiency by controlling local update frequency and batch size. The theoretical analysis demonstrates convergence rates, which facilitate the creation of an adaptive algorithm for adjusting update frequency and batch sizes tailored to heterogeneous workers. However, clients must obtain back-propagation results from the server at each local update. Meanwhile, [8] recommends updating client and server-side models concurrently, utilizing local-loss-based training and auxiliary networks designed specifically for split learning. This parallel training approach effectively reduces latency and eliminates the need for server-to-client communication. The paper includes latency analysis for optimal model partitioning and offers guidelines for model splitting. Specifically, [4] developed a communication and storage-efficient SFL approach. In this method, each client trains a portion of the model and calculates its local loss function using an auxiliary network, leading to reduced communication overhead. Furthermore, the server model is trained based on the sequence of forward propagation results from the clients, ensuring that only one copy of the server model is maintained at any given time. Additionally, [8] suggested a similar framework, albeit with a key difference that each client possesses its separate server model, and these models are aggregated to construct the global server model.
3. **Auxiliary networks**

   Neural network training with back-propagation is hindered by inefficiencies arising from the update locking issue, where layers must await the complete propagation of signals through the network before updating [9]. To address this, [9] proposed Decoupled Greedy Learning (DGL), a more straightforward training approach that relaxes the joint training objective greedily, showing significant effectiveness for CNNs in large-scale image classification. This method optimizes the training objective using auxiliary modules or replay buffers to reduce communication delays caused by waiting for backward propagation. [10] addressed the backward update lock constraint by introducing a model that decouples modules through predictions of future computations within the network graph. These models use local information to predict the outcomes of subgraphs, particularly focusing on error gradients. By using synthetic gradients instead of true backpropagated gradients, subgraphs can update independently and asynchronously, realizing decoupled neural interfaces. A similar approach has been adopted for training in SFL by [4,8]. Indeed, they use an auxiliary model to replace the server model. The mentioned research demonstrates that an auxiliary model with a relatively smaller dimension compared to the server model performs sufficiently well in serving as a replacement.

3. SplitFed Learning Scenario

In this section, we introduce the SFL framework, encompassing both client and server-side models. Additionally, we present the CSE-SFL algorithm designed by [4] to mitigate communication overhead. Accordingly, we split the model as $\tilde{\mathbf{x}} := (\mathbf{x}_C, \mathbf{x}_S)$ where $\mathbf{x}_C$ denotes the client-side model, and $\mathbf{x}_S$ indicates the server-side model. We introduce $\mathbf{x} := (\mathbf{x}_C, \mathbf{x}_A)$ as a client-side model including the auxiliary network where $\mathbf{x}_A$ indicates the model for the auxiliary network.

The client-side non-convex loss function in the SFL setting is given by:

$$F_i^c(\mathbf{x}) \triangleq \mathbb{E}_{\xi \sim D_i}\left[F_i^c(\mathbf{x};\xi)\right] \tag{2}$$

Also, the non-convex loss function in the SFL setting is defined by:

$$F^s(\mathbf{x}_S; \mathbf{z}_f, \mathbf{y}) \triangleq \frac{1}{m}\sum_{i=0}^{m-1} F_i^s(\mathbf{x}_S; \mathbf{z}_{f,i}, \mathbf{y}_i) \tag{3}$$

We denote $\mathbf{z}_{f,i}(\mathbf{x}_C;\xi)$ as the output of the forward propagation of the client i's model, $\mathbf{x}_{C,i}$, on its local random data sample, $\xi \in D_i$, which is intended to be transmitted to the server at specific intervals including the true labels $\mathbf{y}_i$ corresponding to the local random data sample. Note that the sampled data at the client is not shared with the server but the true labels. Similarly, $\mathbf{z}_{b,i}(\mathbf{x}_S; \mathbf{z}_{f,i}, \mathbf{y}_i)$ indicates the backward propagation model of the server for client i. Accordingly, $\hat{\mathbf{z}}_{b,i}(\mathbf{x}_A; \mathbf{z}_{f,i}, \mathbf{y}_i)$ corresponds to the backward propagation results obtained by the auxiliary network. In more detail, the client i performs forward propagation up to the splitting layer and transmits the output of this layer, along with the true labels, to the server. The server then continues forward propagation through to the final layer and computes the loss function. Subsequently, the server performs backward propagation of the error and sends the gradients of its first layer back to the client. We consider $\bar{\mathbf{x}}_C^t$ as the aggregated model at each global round $t \in [T]$ where $[T] = \{0, ..., T-1\}$ and $\bar{\mathbf{x}}_C^t = \frac{1}{m}\sum_i \mathbf{x}_C^t$. Throughout this paper, $[S] = \{0, ..., m-1\}$ identifies the clients' set which is indexed by i. We employ two strategies for client participation. The first strategy entails all clients participating in the learning process. The second strategy involves the server randomly sampling a subset of size n of clients with replacement, $[S_t]$, following a uniform distribution. We assume that D_is are non-i.i.d. The derivative of local loss function of client i in SFL setting with respect to $\mathbf{x}_C$ and $\mathbf{x}_A$ are indicated by $\nabla F_i^c(\mathbf{x}_C)$ and $\nabla F_i^c(\mathbf{x}_A)$ respectively. As for the server-side model, the derivative of the loss function is $\nabla F^s(\mathbf{x}_S)$ which is with respect to $\mathbf{x}_S$. The stochastic gradients of each of the aforementioned gradients will be distinguished by a $\tilde{\nabla}$ sign, e.g., $\tilde{\nabla} F_i^c(\mathbf{x}_C) = \nabla F_i^c(\mathbf{x}_C;\xi)$ where $\xi \sim D_i$ is a random sample from client i dataset. Note that μ_L, and μ are the learning rates of client-side and server-side models respectively. Client i trains $\mathbf{x}_{C,i}$ on its local dataset and renders the forward propagation results, $\mathbf{z}_{f,i}$, to the auxiliary network at each local step k and it receives the $\hat{\mathbf{z}}_{b,i}$ in response. Note that $k \in [K]$ indexes the local steps. Additionally, the client sends the $\mathbf{z}_{f,i}$ to the server at each global round t such that $t \equiv 0 \mod l$ where l is a parameter determining the frequency of this process. We have one server performing the model aggregation at each global round, completing the forward propagation of clients, and updating the server model at specific global rounds. Algorithm 1 illustrates the proposed procedure by [4] in detail.

Algorithm 1 CSE-SFL [4]

```
At Server
Initialize $\mathbf{x}_C^0$, $\mathbf{x}_A^0$ and $\mathbf{x}_S^0$
for $t = 0, 1, ..., T-1$ do
    Sample a subset $S_t$ of $n$ clients out of $m$ clients
    Receive $\mathbf{x}_{C,i}^t, \mathbf{x}_{A,i}^t$ $\forall i \in [S_t]$
    Let $\bar{\mathbf{x}}_C^t \leftarrow \frac{1}{m} \sum_{i \in [S_t]} \mathbf{x}_{C,i}^t$ and $\bar{\mathbf{x}}_A^t \leftarrow \frac{1}{m} \sum_{i \in [S_t]} \mathbf{x}_{A,i}^t$
    Broadcast $\bar{\mathbf{x}}_C^t$ and $\bar{\mathbf{x}}_A^t$ to clients
    if $t \equiv 0 \mod l$, and $t \neq 0$ then
        for each client $i \in [S_t]$ in sequence do
            $\mathbf{z}_{f,i}, \mathbf{y}_i \leftarrow$ Client$(i, \mathbf{z}_f, \mathbf{y})$
            Complete forward propagation with $\mathbf{z}_{f,i}$, and $\mathbf{x}_S^0$
            Compute $\hat{\mathbf{y}}_i$, the prediction of $\mathbf{y}_i$
            Compute loss function $F_i^s(\mathbf{x}_S^0; \mathbf{z}_{f,i}, \mathbf{y}_i)$
            Complete backward-propagation
            Send $\mathbf{z}_{b,i}$ to the client
            Update server model: $\mathbf{x}_S^0 \leftarrow \mathbf{x}_S^0 - \frac{\mu}{m} \nabla F_i^s(\mathbf{x}_S^0; \mathbf{z}_{f,i}, \mathbf{y}_i)$
        end for
    end if
end for
Concatenate $\mathbf{x}_C$ and $\mathbf{x}_S$
At Clients :
for all clients $i \in [S_t]$ in parallel at round $t$ do
    $\mathbf{x}_{C,i}^0, \leftarrow$ Server$(\bar{\mathbf{x}}_C^t)$
    if $t \equiv 0 \mod l$, and $t \neq 0$ then
        $\mathbf{z}_{f,i} \leftarrow$ ForwardPass$(\mathbf{x}_{C,i}^0; \xi)$
        Send $\mathbf{z}_{f,i}$ and $\mathbf{y}_i$ to the server
        $\mathbf{z}_{b,i}^t, \leftarrow$ Server$(\mathbf{z}_b^t)$
        Complete backward-propagation with $\mathbf{z}_{b,i}^t$
        Client update: $\mathbf{x}_{C,i}^1 \leftarrow \mathbf{x}_{C,i}^0 - \mu_L \nabla F_i^c(\mathbf{x}_{C,i}^0)$
        Auxiliary update: $\mathbf{x}_{A,i}^1 \leftarrow \mathbf{x}_{A,i}^0$
        for local step $k = 1, .., K-1$ do
            Compute forward propagation with $\mathbf{x}_{C,i}^k$ and $\mathbf{x}_A^t$
            Compute local loss $F_i^c(\mathbf{x}_i^k; \xi^k)$
            Client update: $\mathbf{x}_{C,i}^{k+1} \leftarrow \mathbf{x}_{C,i}^k - \mu_L \nabla F_i^c(\mathbf{x}_{C,i}^k)$
            Auxiliary update: $\mathbf{x}_{A,i}^{k+1} \leftarrow \mathbf{x}_{A,i}^k - \mu_L \nabla F_i^c(\mathbf{x}_{A,i}^k)$
        end for
    else
        for local step $k = 0, .., K-1$ do
            Compute forward propagation with $\mathbf{x}_{C,i}^k$ and $\mathbf{x}_A^t$
            Compute local loss $F_i^c(\mathbf{x}_i^k; \xi^k)$
            Client update: $\mathbf{x}_{C,i}^{k+1} \leftarrow \mathbf{x}_{C,i}^k - \mu_L \nabla F_i^c(\mathbf{x}_{C,i}^k)$
            Auxiliary update: $\mathbf{x}_{A,i}^{k+1} \leftarrow \mathbf{x}_{A,i}^k - \mu_L \nabla F_i^c(\mathbf{x}_{A,i}^k)$
        end for
    end if
    Return $\mathbf{x}_{C,i}^K$ to the server
end for
```

4. Convergence rate analysis

The following assumptions for the convergence rate evaluation have been made:

Assumption 1. *(L-Lipschitz continuous gradient) Both client and server-side models are $L-$smooth non-convex functions, i.e., there is a constant $L > 0$ such that $\forall \mathbf{x}_C, \mathbf{y}_C \in \mathbb{R}^{d_c}$, and $\forall \mathbf{x}_S, \mathbf{y}_S \in \mathbb{R}^{d_s}$:*

$$\|\nabla F^c(\mathbf{x}_C) - \nabla F^c(\mathbf{y}_C)\| \leq L\|\mathbf{x}_C - \mathbf{y}_C\| \quad \textit{and} \quad \|\nabla F^s(\mathbf{x}_S) - \nabla F^s(\mathbf{y}_S)\| \leq L\|\mathbf{x}_S - \mathbf{y}_S\|$$

Assumption 2. *(Unbiased local gradient estimator) We assume that $\forall i \in [S]$,*

$$\mathbb{E}_{\xi \in D_i}\left[\nabla F_i^c(\mathbf{x}_C; \xi)\right] = \nabla F_i^c(\mathbf{x}_C)$$

that is the local gradient estimator of the client-side model is unbiased. The expectation is over all the local datasets of the client. Note that we have a similar assumption for the server-side model as follows $\forall i \in [S]$*:*

$$\mathbb{E}_{\xi \in D_i}\left[\nabla F_i^s(\mathbf{x}_S; \mathbf{z}_{f,i}(\mathbf{x}_C; \xi))\right] = \nabla F_i^s(\mathbf{x}_S)$$

Assumption 3. *(Bounded local and global variance) We have bounded variance of the stochastic gradients locally and globally for both server-side and client-side models, i.e., there exist positive constants* σ_L *and* σ_G *such that*

$$\mathbb{E}\left[\|\nabla F_i^c(\mathbf{x}_C;\xi) - \nabla F_i^c(\mathbf{x}_C)^2\|\right] \le \sigma_L^2 \quad \text{and} \quad \mathbb{E}\left[\|\nabla F_i^c(\mathbf{x}_C) - \nabla F^c(\mathbf{x}_C)^2\|\right] \le \sigma_G^2$$
$$\mathbb{E}\left[\|\nabla F_i^s(\mathbf{x}_S;\xi) - \nabla F_i^s(\mathbf{x}_S)^2\|\right] \le \sigma_L^2 \quad \text{and} \quad \mathbb{E}\left[\|\nabla F_i^s(\mathbf{x}_S) - \nabla F^s(\mathbf{x}_S)^2\|\right] \le \sigma_G^2$$

Assumptions 1, 2, and 3 are natural assumptions applied in non-convex optimization and FL, e.g., see [7,11–15]. Figure 1 gives an overview of the communication and storage of efficient federated split learning (CSE-FSL) algorithm in an illustrative way.

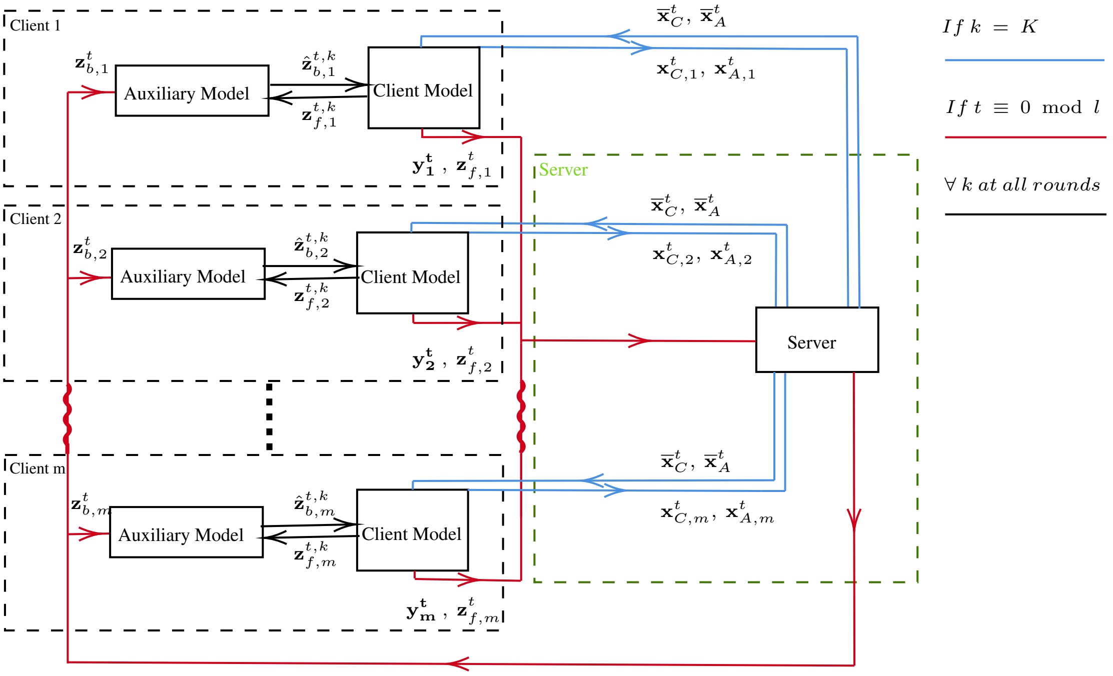
Client 1
$\mathbf{z}_{b,1}^t$
Auxiliary Model
$\hat{\mathbf{z}}_{b,1}^{t,k}$
$\mathbf{z}_{f,1}^{t,k}$
Client Model
$\mathbf{y}_1^t$, $\mathbf{z}_{f,1}^t$
Client 2
$\mathbf{z}_{b,2}^t$
Auxiliary Model
$\hat{\mathbf{z}}_{b,2}^{t,k}$
$\mathbf{z}_{f,2}^{t,k}$
Client Model
$\mathbf{y}_2^t$, $\mathbf{z}_{f,2}^t$
Client m
$\mathbf{z}_{b,m}^t$
Auxiliary Model
$\hat{\mathbf{z}}_{b,m}^{t,k}$
$\mathbf{z}_{f,m}^{t,k}$
Client Model
$\mathbf{y}_m^t$, $\mathbf{z}_{f,m}^t$
$\bar{\mathbf{x}}_C^t$, $\bar{\mathbf{x}}_A^t$
$\mathbf{x}_{C,1}^t$, $\mathbf{x}_{A,1}^t$
$\bar{\mathbf{x}}_C^t$, $\bar{\mathbf{x}}_A^t$
$\mathbf{x}_{C,2}^t$, $\mathbf{x}_{A,2}^t$
$\bar{\mathbf{x}}_C^t$, $\bar{\mathbf{x}}_A^t$
$\mathbf{x}_{C,m}^t$, $\mathbf{x}_{A,m}^t$
Server
Server
$If\ k = K$
$If\ t \equiv 0 \mod l$
$\forall\ k\ at\ all\ rounds$

Figure 1. CSE-FSL pipeline

4.1. Client-Side Model Convergence

We examine the convergence rate when $t \equiv 0 \mod l$ because it is during these rounds that the server-side model is also updated. This will let us study the impact of l on the convergence rate and communication overhead.

Theorem 1. *Under Assumptions 1, 2, 3, and full participation of clients, if* $\mu_L \le \frac{1}{lLK2^{1.15l+1.85}}$*, and* $t \equiv 0 \mod l$*, in Algorithm 1, the convergence rate of client model of Algorithm 1 satisfies:*

$$\min_{t\in[T]} \mathbb{E}\left[\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\|\right] \le \frac{2\left(F^c(\bar{\mathbf{x}}_C^0) - F^c(\bar{\mathbf{x}}_C^*)\right)}{(1-\Gamma\lambda_2)\mu_L KT} + \Phi_1 + \frac{\Gamma\lambda_1}{1-\Gamma\lambda_2}$$

Where,

$$\Phi_1 = \frac{5K\mu_L^2L^2(\sigma_L^2 + 6K\sigma_G^2))}{1-\Gamma\lambda_2} +$$
$$\frac{4L\mu_L + 4\mu_L^2KL^2(l-1)}{1-\Gamma\lambda_2}\Big(\big(Kl + 5L^2K^2l\mu_L^2\big)\sigma_L^2 + \big(Kl + 30L^2K^3l\mu_L^2\big)\sigma_G^2\Big)$$
$$\lambda_1 = B\sum_{j=0}^{l-1}\frac{A^j-1}{A-1}, \quad \lambda_2 = \frac{A^l-1}{A-1},$$
$$B = 8L^2\mu_L^2\Big(\big(K^2 + 5L^2K^3\mu_L^2\big)\sigma_L^2 + \big(K^2 + 30L^2K^4\mu_L^2\big)\sigma_G^2\Big),$$
$$A = 8L^2\mu_L^2(K^2 + 30L^2K^3\mu_L^2) + 2,$$
$$\Gamma = 4\Big(L\mu_L + \mu_L^2KL^2(l-1)\Big)\Big(K + 30L^2K^2\mu_L^2\Big) + 30K^2\mu_L^2L^2\Big), \text{ and}$$
$$\bar{\mathbf{x}}_C^* = \operatorname*{argmin}_{\bar{\mathbf{x}}_C^t, t\in[T]} \mathbb{E}\Big[\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\|\Big]$$

Corollary 1. *Let $\mu_L \le \frac{1}{lLK2^{1.15l+1.85}\sqrt{T}}$. Then, the convergence rate of the client-side model in Algorithm 1 is*

$$\min_{t\in[T]} \mathbb{E}\Big[\|\nabla f^c(\bar{\mathbf{x}}_C^t)^2\|\Big] \le \mathcal{O}\Big(\frac{l}{\sqrt{T}} + \frac{1}{T\sqrt{T}}\Big). \tag{4}$$

Theorem 2. *Under Assumptions 1, 2, 3, and partial participation of clients due to strategy one, if $\mu_L \le \frac{1}{lLK2^{1.15l+1.85}}$, and $t \equiv 0 \mod l$, in Algorithm 1, the convergence rate of client model of Algorithm 1 satisfies:*

$$\min_{t\in[T]} \mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| \le \frac{2\big(F^c(\bar{\mathbf{x}}_C^0) - F^c(\bar{\mathbf{x}}_C^*)\big)}{\mu_L KT} +$$
$$\Big(5K\mu_L^2L^2 + 4\mu_L^2L^2\Big(K^2l^2 + 5L^2K^3l^2\mu_L^2\Big) + L\mu_L\Big(\frac{1}{n} + 15K^2L^2\mu_L^2\Big)\Big)\sigma_L^2 +$$
$$\Big(30K^2\mu_L^2L^2 + L\mu_L\Big(90K^3L^2\mu_L^2 + 3K\Big) + 4\mu_L^2L^2\Big(K^2l^2 + 30L^2K^4l^2\mu_L^2\Big)\Big)\sigma_G^2 +$$
$$\frac{\Gamma'\lambda_1}{1-\Gamma'\lambda_2}$$

Where

$$\Gamma' = 4\mu_L^2L^2\Big(K^2l + 30L^2K^3l\mu_L^2\Big) + \frac{L\mu_L}{l}\Big(90lK^3L^2\mu_L^2 + 3K\Big) + 30K^2\mu_L^2L^2$$

and,

$$\bar{\mathbf{x}}_C^* = \operatorname*{argmin}_{\bar{\mathbf{x}}_C^t, t\in[T]} \mathbb{E}\Big[\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\|\Big]$$

Corollary 2. *Let $\mu_L \le \frac{1}{lLK2^{1.15l+1.85}\sqrt{T}}$. Then, the convergence rate of the client-side model in Algorithm 1 is*

$$\min_{t\in[T]} \mathbb{E}\Big[\|\nabla f^c(\bar{\mathbf{x}}_C^t)^2\|\Big] \le \mathcal{O}\Big(\frac{l}{\sqrt{T}} + \frac{1}{T\sqrt{T}}\Big). \tag{5}$$

4.2. Server-Side Model Convergence

Theorem 3. *Under Assumptions 1, 2, 3, and full participation of clients, if $\mu \leq \frac{1}{2L}$, and $t \equiv 0 \mod l$, the convergence rate of the server model of Algorithm 1 satisfies:*

$$\min_{t\in[T]} \mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| \leq \frac{2l\left(F^s(\mathbf{x}_S^0) - F^s(\mathbf{x}_S^*)\right)}{\mu(2m-3)T} + \frac{L\mu m^2}{2m-3}\left(9.2\sigma_L^2 + 13.2\sigma_G^2\right)$$

Where $\mathbf{x}_S^* = \operatorname{argmin}_{\mathbf{x}_S^t, t\in[T]} \mathbb{E}\|\nabla F^s(\mathbf{x}_S^t, \mathbf{z}_f^t)^2\|$.

Corollary 3. *Let $\mu \leq \frac{1}{2L\sqrt{T}}$, then the convergence rate of the server-side model is:*

$$\min_{t\in[T]} \mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| \leq \mathcal{O}\left(\frac{l}{\sqrt{T}}\right)$$

Theorem 4. *Under Assumptions 1, 2, 3, and partial participation of clients due to strategy one, if $\mu \leq \frac{1}{8L^2m^2}$, and $t \equiv 0 \mod l$, the convergence rate of the server model of Algorithm 1 satisfies:*

$$\min_{t\in[T]} \mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| \leq \frac{l\left(F^s(\mathbf{x}_S^0) - F^s(\mathbf{x}_S^*)\right)}{\mu(m-2)T} + \frac{L\mu m^2}{m-2}\left(7\sigma_L^2 + 7\sigma_G^2\right)$$

Where $\mathbf{x}_S^* = \operatorname{argmin}_{\mathbf{x}_S^t} \mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2$.

$\|$

Corollary 4. *Let $\mu \leq \frac{1}{L^2m^2\sqrt{T}}$, then the convergence rate of the server-side model is:*

$$\min_{t\in[T]} \mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| \leq \mathcal{O}\left(\frac{l}{\sqrt{T}}\right)$$

5. Discussion and Conclusions

In this paper, we proposed theoretical convergence proofs for the state-of-the-art SplitFed Learning algorithm, CSE-FSL, which is designed to improve the convergence rates of both client-side and server-side models leveraging parallelism power of Federated Learning (FL) and reduce the storage at the server by keeping one copy at a time policy. Our approach leverages several key assumptions that are conventional in FL to underpin the theoretical foundations for CSE-FSL convergence. We prove the convergence for the cases where we have non-i.i.d. datasets, and non-convex loss functions given full and partial client participation scenarios.

5.1. Summary of Contributions

- **Convergence Analysis**: We clearly formulated the CSE-FSL algorithm developed by [4]. We conducted a comprehensive convergence rate analysis under both full and partial client participation scenarios given the non-i.i.d. dataset and non-convex loss function. The convergence guarantees are derived under several assumptions, including L-smoothness of the objective functions, unbiased gradient estimators, and bounded gradient variances which are natural in conventional FL convergence analysis.
- **Key Results**:
  - *Client-Side Model*: We demonstrated that, under full client participation, the client-side model converges with a rate of $\mathcal{O}\left(\frac{l}{\sqrt{T}} + \frac{1}{T\sqrt{T}}\right)$. This result highlights the effectiveness of the algorithm in achieving linear convergence rates while accommodating the federated

setting's constraints and sequential update of the server model. An increase in l, causes a longer convergence time which is obvious as it means the server model will be updated after more global rounds.

- *Server-Side Model*: For the server-side model, we established convergence rates of $\mathcal{O}\left(\frac{l}{\sqrt{T}}\right)$ under both full and partial client participation scenarios. This result underscores the robustness of the algorithm in ensuring effective learning even when clients participate partially. This also demonstrates that the number of clients and their local steps are not effective in speeding up the convergence in contrast to FL settings.

5.2. Implications

Our findings underscore the importance of efficient communication and gradient estimation (auxiliary networks) techniques in SplitFed Learning (SFL). The derived convergence rates demonstrate that the CSE-FSL algorithm achieves a balance between computational efficiency and convergence performance, making it a viable solution for practical federated learning applications where the privacy of clients is of high importance.

The theoretical guarantees provided by our convergence analysis offer valuable insights into how the algorithm performs under various conditions, thus guiding practitioners in optimizing federated learning systems. Future work could extend these results to explore more complex scenarios and refine the algorithm further for enhanced performance in real-world applications. For example, considering stragglers, elimination of label sharing by clients, and determining the optimal cut layer seem to be promising avenues for further research.

In summary, the CSE-FSL algorithm represents a significant advancement in FL, providing a robust framework for effective model training leveraging the parallelism power of FL, auxiliary networks, and sequential updates of the server-side model which helps reduce storage on the server side. It recovers the linear convergence speed of FL while providing more privacy by only forward-propagation and label transition between clients and servers instead of trained parameters.

Appendix A. Proofs

Appendix A.1. Client-Side Model Convergence

We examine the convergence rate when $t \equiv 0 \mod l$ because it is during these rounds that the server-side model is also updated. This analysis allows us to investigate the influence of l on both the convergence rate and communication overhead.

Theorem A1. *Under Assumptions 1, 2, 3, and full participation of clients, if $\mu_L \leq \frac{1}{lLK2^{1.15l+1.85}}$, and $t \equiv 0 \mod l$, in Algorithm 1, the convergence rate of client model of Algorithm 1 satisfies:*

$$\min_{t\in[T]} \mathbb{E}\left[\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\|\right] \leq \frac{2\left(F^c(\bar{\mathbf{x}}_C^0) - F^c(\bar{\mathbf{x}}_C^*)\right)}{(1-\Gamma\lambda_2)\mu_L KT} + \Phi_1 + \frac{\Gamma\lambda_1}{1-\Gamma\lambda_2}$$

Where,

$$\Phi_1 = \frac{5K\mu_L^2 L^2(\sigma_L^2 + 6K\sigma_G^2))}{1-\Gamma\lambda_2} +$$
$$\frac{4L\mu_L + 4\mu_L^2 KL^2(l-1)}{1-\Gamma\lambda_2}\left(\left(Kl + 5L^2K^2l\mu_L^2\right)\sigma_L^2 + \left(Kl + 30L^2K^3l\mu_L^2\right)\sigma_G^2\right)$$
$$\lambda_1 = B\sum_{j=0}^{l-1}\frac{A^j - 1}{A-1}, \quad \lambda_2 = \frac{A^l - 1}{A-1},$$
$$B = 8L^2\mu_L^2\left(\left(K^2 + 5L^2K^3\mu_L^2\right)\sigma_L^2 + \left(K^2 + 30L^2K^4\mu_L^2\right)\sigma_G^2\right),$$
$$A = 8L^2\mu_L^2(K^2 + 30L^2K^3\mu_L^2) + 2,$$
$$\Gamma = 4\left(L\mu_L + \mu_L^2 KL^2(l-1)\right)\left(K + 30L^2K^2\mu_L^2\right) + 30K^2\mu_L^2L^2\Big), \text{ and}$$
$$\bar{\mathbf{x}}_C^* = \operatorname*{argmin}_{\bar{\mathbf{x}}_C^t, t\in[T]} \mathbb{E}\left[\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\|\right]$$

Corollary A1. *Let* $\mu_L \le \frac{1}{lLK2^{1.15l+1.85}\sqrt{T}}$. *Then, the convergence rate of the client-side model in Algorithm 1 is*

$$\min_{t\in[T]} \mathbb{E}\left[\|\nabla f^c(\bar{\mathbf{x}}_C^t)^2\|\right] \le \mathcal{O}\left(\frac{l}{\sqrt{T}} + \frac{1}{T\sqrt{T}}\right). \tag{A1}$$

Proof. In this proof, all the gradients are w.r.t. $\mathbf{x}_C$. Due to Assumption 1, for any $\bar{\mathbf{x}}_C^{t+l}$ and $\bar{\mathbf{x}}_C^t$ such that $t \in [T]$, we can write:

$$F^c(\bar{\mathbf{x}}_C^{t+l}) \le F^c(\bar{\mathbf{x}}_C^t) + \nabla F^c(\bar{\mathbf{x}}_C^t)^\top(\bar{\mathbf{x}}_C^{t+l} - \bar{\mathbf{x}}_C^t) + \frac{L}{2}\|\bar{\mathbf{x}}_C^{t+l} - \bar{\mathbf{x}}_C^{t}{}^2\| \tag{A2}$$

Particularly, we consider the case when $t \equiv 0 \mod l$ from now on.

Also, note the global aggregation and client update rule in the Algorithm 1,

$$\bar{\mathbf{x}}_C^{t+l} = \frac{1}{m}\sum_{i=0}^{m-1}\mathbf{x}_{C,i}^{t+l} = \frac{1}{m}\sum_{i=0}^{m-1}\left(\mathbf{x}_{C,i}^t - \mu_L\sum_{j=0}^{l-1}\sum_{k=0}^{K-1}\tilde{\nabla}F_i^c(\mathbf{x}_{C,i}^{t+j,k})\right) =$$
$$\bar{\mathbf{x}}_C^t - \frac{\mu_L}{m}\sum_{i=0}^{m-1}\sum_{j=0}^{l-1}\sum_{k=0}^{K-1}\tilde{\nabla}F_i^c(\mathbf{x}_{C,i}^{t+j,k}) \tag{A3}$$

Thus,

$$\bar{\mathbf{x}}_C^{t+l} - \bar{\mathbf{x}}_C^t = -\frac{\mu_L}{m}\sum_{i=0}^{m-1}\sum_{j=0}^{l-1}\sum_{k=0}^{K-1}\tilde{\nabla}F_i^c(\mathbf{x}_{C,i}^{t+j,k}) \tag{A4}$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2024 doi:10.20944/preprints202409.0335.v1

Taking expectation of $F^c(\mathbf{x}_C^{t+1})$ with respect to randomness at round $t+l-1$, i.e., $¸^{[t+l-1]} \triangleq [\xi_i^\tau]_{i\in[N],\tau\in[t+l-1]}$, and plugging A3 into A2 note that:

$$\mathbb{E}\left[F^c(\bar{\mathbf{x}}_C^{t+l})\right] \leq F^c(\bar{\mathbf{x}}_C^t) + \left\langle \nabla F^c(\bar{\mathbf{x}}_C^t), \mathbb{E}\left[\bar{\mathbf{x}}_C^{t+l} - \bar{\mathbf{x}}_C^t\right]\right\rangle + \frac{L}{2}\mathbb{E}\|\bar{\mathbf{x}}_C^{t+l} - \bar{\mathbf{x}}_C^{t}{}^2\|$$

$$\mathbb{E}\left[F^c(\bar{\mathbf{x}}_C^{t+l})\right] \leq F^c(\bar{\mathbf{x}}_C^t) + \mu_L \underbrace{\left\langle \nabla F^c(\bar{\mathbf{x}}_C^t), \mathbb{E}\left[\frac{-1}{m}\sum_{i=0}^{m-1}\sum_{j=0}^{l-1}\sum_{k=0}^{K-1}(\tilde{\nabla} F_i^c(\mathbf{x}_{C,i}^{t+j,k}) - \nabla F_i^c(\bar{\mathbf{x}}_C^{t+j}))\right]\right\rangle}_{A_1}$$

$$\underbrace{-\mu_L\left\langle \nabla F^c(\bar{\mathbf{x}}_C^t), \mathbb{E}\left[\frac{K}{m}\sum_{i=0}^{m-1}\sum_{j=0}^{l-1}\nabla F_i^c(\bar{\mathbf{x}}_C^{t+j})\right]\right\rangle}_{A_2} + \frac{L\mu_L^2}{2}\underbrace{\mathbb{E}\left[\|\frac{1}{m}\sum_{i=0}^{m-1}\sum_{j=0}^{l-1}\sum_{k=0}^{K-1}\tilde{\nabla} F_i^c(\mathbf{x}_{C,i}^{t+j,k})^2\|\right]}_{A_3} \tag{A5}$$

We bound the term A_1 as follows:

$$\begin{aligned}
A_1 &= \left\langle \nabla F^c(\bar{\mathbf{x}}_C^t), \mathbb{E}\left[\frac{-1}{m}\sum_{i=0}^{m-1}\sum_{j=0}^{l-1}\sum_{k=0}^{K-1}(\tilde{\nabla} F_i^c(\mathbf{x}_{C,i}^{t+j,k}) - \nabla F_i^c(\bar{\mathbf{x}}_C^{t+j}))\right]\right\rangle \\
&= \left\langle \nabla F^c(\bar{\mathbf{x}}_C^t), \mathbb{E}\underbrace{\left[\frac{-1}{m}\sum_{i=0}^{m-1}\sum_{j=0}^{l-1}\sum_{k=0}^{K-1}(\nabla F_i^c(\mathbf{x}_{C,i}^{t+j,k}) - \nabla F_i^c(\bar{\mathbf{x}}_C^{t+j}))\right]}_{y_1}\right\rangle \\
&\overset{(a_1)}{=} \frac{K}{2}\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| + \frac{1}{2Km^2}\mathbb{E}\|\sum_{i=0}^{m-1}\sum_{j=0}^{l-1}\sum_{k=0}^{K-1}(\nabla F_i^c(\mathbf{x}_{C,i}^{t+j,k}) - \nabla F_i^c(\bar{\mathbf{x}}_C^{t+j}))^2\| \\
&- \frac{1}{2Km^2}\mathbb{E}\|\sum_{i=0}^{m-1}\sum_{j=0}^{l-1}\sum_{k=0}^{K-1}(\nabla F_i^c(\mathbf{x}_{C,i}^{t+j,k}) - \nabla F_i^c(\bar{\mathbf{x}}_C^{t+j})) + K\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| \\
&\overset{(a_2)}{\leq} \frac{K}{2}\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2 + \frac{1}{2Km^2}\mathbb{E}\|\sum_{i=0}^{m-1}\sum_{j=0}^{l-1}\sum_{k=0}^{K-1}(\nabla F_i^c(\mathbf{x}_{C,i}^{t+j,k}) - \nabla F_i^c(\bar{\mathbf{x}}_C^{t+j}))^2\| \\
&\overset{(a_3)}{\leq} \frac{K}{2}\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| + \frac{lL^2}{2m}\sum_{i=0}^{m-1}\sum_{j=0}^{l-1}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla F_i^c(\mathbf{x}_{C,i}^{t+j,k}) - \nabla F_i^c(\bar{\mathbf{x}}_C^{t+j})^2\|\right] \\
&\overset{(a_4)}{\leq} \frac{K}{2}\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| + \frac{lL^2}{2m}\sum_{i=0}^{m-1}\sum_{j=0}^{l-1}\sum_{k=0}^{K-1}\underbrace{\mathbb{E}\left[\|\mathbf{x}_{C,i}^{t+j,k} - \bar{\mathbf{x}}_C^{t+j}{}^2\|\right]}_{y_2}
\end{aligned} \tag{A6}$$

$$\text{We have } \langle a, b\rangle = \frac{1}{2}(\|a^2\| + \|b^2\| - \|a - b^2\|) \text{ for any two vectors } a \text{ and } b. \tag{A7}$$

Thus, if we put $a = \sqrt{K}\nabla F^c(\bar{\mathbf{x}}_C^t)$, and $b = \frac{1}{\sqrt{K}}y_1$, it yields equality (a_1). Inequality (a_2) follows from eliminating a strictly negative term. Now, due to $\mathbb{E}\|\sum_i^n z_i{}^2\| \leq n\sum_i \mathbb{E}\left[\|z_i{}^2\|\right]$ for any random variables z_i, inequality (a_3) holds. (A8)

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2024 doi:10.20944/preprints202409.0335.v1

The inequality a_4 follows from Assumption 1. There is an upper bound for the term y_2 provided by [13]. To preserve the integrity of the work, we include it here as well.

$$
\begin{aligned}
&\mathbb{E}\left[\|\mathbf{x}_{C,i}^{t+j,k} - \bar{\mathbf{x}}_C^{t+j^2}\|\right] = \mathbb{E}\left[\|\mathbf{x}_{C,i}^{t+j,k-1} - \bar{\mathbf{x}}_C^{t+j} - \mu_L \tilde{\nabla} F_i^c(\mathbf{x}_{C,i}^{t+j,k-1})^2\|\right] \\
&= \mathbb{E}\left\|\mathbf{x}_{C,i}^{t+j,k-1} - \bar{\mathbf{x}}_C^{t+j} - \mu_L\Big(\tilde{\nabla} F_i^c(\mathbf{x}_{C,i}^{t+j,k-1}) - \nabla F_i^c(\mathbf{x}_{C,i}^{t+j,k-1}) + \nabla F_i^c(\mathbf{x}_{C,i}^{t+j,k-1}) \right. \\
&\left. - \nabla F_i^c(\bar{\mathbf{x}}_C^{t+j}) + \nabla F_i^c(\bar{\mathbf{x}}_C^{t+j}) - \nabla F^c(\bar{\mathbf{x}}_C^{t+j}) + \nabla F^c(\bar{\mathbf{x}}_C^{t+j})\Big)\right\|^2 \\
&\overset{(a_5)}{\leq} (1+\frac{1}{2K-1})\mathbb{E}\|\mathbf{x}_{C,i}^{t+j,k-1} - \bar{\mathbf{x}}_C^{t+j^2}\| + \mathbb{E}\|\mu_L(\tilde{\nabla} F_i^c(\mathbf{x}_{C,i}^{t+j,k-1}) - \nabla F_i^c(\mathbf{x}_{C,i}^{t+j,k-1}))^2\| + \\
&6K\mathbb{E}\|\mu_L(\nabla F_i^c(\mathbf{x}_{C,i}^{t+j,k-1}) - \nabla F_i^c(\bar{\mathbf{x}}_C^{t+j}))^2\| + 6K\mathbb{E}\|\mu_L(\nabla F_i^c(\bar{\mathbf{x}}_C^{t+j}) - \nabla F^c(\bar{\mathbf{x}}_C^{t+j}))^2\| \\
&+ 6K\mathbb{E}\|\mu_L \nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\| \\
&\overset{(a_6)}{\leq} (1+\frac{1}{K-1})\mathbb{E}\|\mathbf{x}_{C,i}^{t+j,k-1} - \bar{\mathbf{x}}_C^{t+j^2}\| + \mu_L^2\sigma_L^2 + 6K\mu_L^2\sigma_G^2 + 6K\mu_L^2\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\|
\end{aligned}
\tag{A9}
$$

The inequality (a_5) follows from the fact that $\mathbb{E}\|\sum_i {z_i}^2\| \leq \mathbb{E}\left[\sum_i \|{z_i}^2\|\right]$ holds true for independent random variables z_i with zero mean.

(A10)

The term (a_6) is due to Assumption 3. Finally, by unrolling recursion and some simplification, we have:

$$
\frac{1}{m}\sum_{i=0}^{m-1} \mathbb{E}\|\mathbf{x}_{C,i}^{t+j,k} - \bar{\mathbf{x}}_C^{t+j^2}\| \leq 5K\mu_L^2(\sigma_L^2 + 6K\sigma_G^2) + 30K^2\mu_L^2\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\| \tag{A11}
$$

Now, we continue by substituting A11 into A6,

$$
\leq \frac{K}{2}\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| + \frac{5K^2\mu_L^2 lL^2}{2}(\sigma_L^2 + 6K\sigma_G^2) + 15K^3\mu_L^2 lL^2 \sum_{j=0}^{l-1}\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\| \tag{A12}
$$

The above inequality, A12, is an upper bound for the term A_1. We continue with bounding A_2 as follows.

$$
\begin{aligned}
A_2 &= -\mu_L\left\langle \nabla F^c(\bar{\mathbf{x}}_C^t), \mathbb{E}\left[\frac{K}{m}\sum_{i=0}^{m-1}\sum_{j=0}^{l-1}\nabla F_i^c(\bar{\mathbf{x}}_C^{t+j})\right]\right\rangle \\
&\overset{(a_7)}{=} -\mu_L K\left\langle \nabla F^c(\bar{\mathbf{x}}_C^t), \mathbb{E}\left[\sum_{j=0}^{l-1}\nabla F^c(\bar{\mathbf{x}}_C^{t+j})\right]\right\rangle \\
&= -\mu_L K\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| - \mu_L K\left\langle \nabla F^c(\bar{\mathbf{x}}_C^t), \mathbb{E}\left[\sum_{j=1}^{l-1}\nabla F^c(\bar{\mathbf{x}}_C^{t+j})\right]\right\rangle \\
&\overset{(a_8)}{\leq} -\mu_L K\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| - \mu_L K\left\langle \nabla F^c(\bar{\mathbf{x}}_C^t), \mathbb{E}\left[(l-1)\nabla F^c(\bar{\mathbf{x}}_C^{t+j^*})\right]\right\rangle
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(a_9)}{=} -\frac{\mu_L K(l+1)}{2}\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| \underbrace{-\frac{\mu_L K(l-1)}{2}\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j^*})^2\|}_{y_3} \\
&+\frac{\mu_L K(l-1)}{2}\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j^*})-\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| \\
&\overset{(a_{10})}{\leq} -\frac{\mu_L K(l+1)}{2}\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| + \frac{\mu_L K(l-1)}{2}\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j^*})-\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| \\
&\overset{(a_{11})}{\leq} -\frac{\mu_L K(l+1)}{2}\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| + \frac{\mu_L K L^2(l-1)}{2}\mathbb{E}\|\bar{\mathbf{x}}_C^{t+j^*}-\bar{\mathbf{x}}_C^{t}{}^2\| \\
&\overset{(a_{12})}{\leq} -\frac{\mu_L K(l+1)}{2}\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| + \frac{\mu_L^3 K L^2(l-1)}{2}\underbrace{\mathbb{E}\|\frac{1}{m}\sum_{i=0}^{m-1}\sum_{j=0}^{j^*}\sum_{k=0}^{K-1}\tilde{\nabla} F_i^c(\mathbf{x}_{C,i}^{t+j,k})^2\|}_{y_4} \qquad \text{(A13)} \\
&\overset{(a_{13})}{\leq} -\frac{\mu_L K(l+1)}{2}\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| + \frac{\mu_L^3 K L^2(l-1)}{2}\underbrace{\mathbb{E}\|\frac{1}{m}\sum_{i=0}^{m-1}\sum_{j=0}^{l-1}\sum_{k=0}^{K-1}\tilde{\nabla} F_i^c(\mathbf{x}_{C,i}^{t+j,k})^2\|}_{A_3} \qquad \text{(A14)}
\end{aligned}
$$

The equality (a_7) follows from definition of the global aggregation in Algorithm 1. In inequality (a_8), we assume there exists a j^* such that $j^* = \text{argmin}_{1\leq j\leq l-1}\left\langle \nabla F^c(\bar{\mathbf{x}}_C^t), \mathbb{E}\left[\nabla F^c(\bar{\mathbf{x}}_C^{t+j})\right]\right\rangle$. The equality (a_9) follows from A7, where $a = \nabla F^c(\bar{\mathbf{x}}_C^t)$ and $b = \nabla F^c(\bar{\mathbf{x}}_C^{t+j^*})$. The inequality (a_{10}) is due to the fact that the term y_3 is negative. Thus, it can be eliminated safely. Due to Assumption 1, we have inequality (a_{11}). The inequality (a_{12}) is due to equation A3. For the term y_4 in A13, there is an upper bound when $j^* = l-1$ due to A15. Hence, with $j^* = l-1$, inequality (a_{13}) is achieved. We proceed with bounding A_3 as follows.

$$
\begin{aligned}
&\mathbb{E}\|\frac{1}{m}\sum_{i=0}^{m-1}\sum_{j=0}^{l-1}\sum_{k=0}^{K-1}\tilde{\nabla} F_i^c(\mathbf{x}_{C,i}^{t+j,k})^2\| \\
&= \frac{1}{m^2}\mathbb{E}\left\|\sum_{i=0}^{m-1}\sum_{j=0}^{l-1}\sum_{k=0}^{K-1}\left(\tilde{\nabla} F_i^c(\mathbf{x}_{C,i}^{t+j,k}) - \nabla F_i^c(\mathbf{x}_{C,i}^{t+j,k}) + \nabla F_i^c(\mathbf{x}_{C,i}^{t+j,k}) - \nabla F_i^c(\bar{\mathbf{x}}_C^{t+j})\right.\right. \\
&\left.\left.+\nabla F_i^c(\bar{\mathbf{x}}_C^{t+j}) - \nabla F^c(\bar{\mathbf{x}}_C^{t+j}) + \nabla F^c(\bar{\mathbf{x}}_C^{t+j})\right)\right\|^2 \\
&\overset{(a_{14})}{\leq} \frac{4Kl}{m}\sum_{i=0}^{m-1}\sum_{j=0}^{l-1}\sum_{k=0}^{K-1}\left(\mathbb{E}\|\tilde{\nabla} F_i^c(\mathbf{x}_{C,i}^{t+j,k}) - \nabla F_i^c(\mathbf{x}_{C,i}^{t+j,k})^2\| + \mathbb{E}\|\nabla F_i^c(\mathbf{x}_{C,i}^{t+j,k}) - \nabla F_i^c(\bar{\mathbf{x}}_C^{t+j})^2\| + \right. \\
&\left.\mathbb{E}\|\nabla F_i^c(\bar{\mathbf{x}}_C^{t+j}) - \nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\| + \mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\|\right) \\
&\overset{(a_{15})}{\leq} \frac{4Kl}{m}\sum_{i=0}^{m-1}\sum_{j=0}^{l-1}\sum_{k=0}^{K-1}\left(\sigma_L^2 + L^2\mathbb{E}\|\mathbf{x}_{C,i}^{t+j,k} - \bar{\mathbf{x}}_C^{t+j}{}^2\| + \sigma_G^2 + \mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\|\right) \\
&\overset{(a_{16})}{\leq} 4\left(K^2l^2 + 5L^2K^3l^2\mu_L^2\right)\sigma_L^2 + 4\left(K^2l^2 + 30L^2K^4l^2\mu_L^2\right)\sigma_G^2 + 4\left(K^2l + \right. \\
&\left.30L^2K^3l\mu_L^2\right)\sum_{j=0}^{l-1}\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\| \qquad \text{(A15)}
\end{aligned}
$$

The inequality (a_{14}) holds due to A8, and inequality (a_{15}) follows from Assumptions 2 and A3. Due to the bound on client drift, A11, note the inequality (a_{16}). Substituting A12, A14, and A15 into A5, observe that:

$$\mathbb{E}\Big[F^c(\bar{\mathbf{x}}_C^{t+l})\Big] \leq F^c(\bar{\mathbf{x}}_C^t)$$
$$+\mu_L\left(\frac{K}{2}\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| + \frac{5K^2\mu_L^2 lL^2}{2}(\sigma_L^2+6K\sigma_G^2) + 15K^3\mu_L^2 lL^2\sum_{j=0}^{l-1}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\|\right)$$
$$-\frac{\mu_L K(l+1)}{2}\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| + 2\Big(L\mu_L^2+\mu_L^3KL^2(l-1)\Big)\bigg(\Big(K^2l^2+5L^2K^3l^2\mu_L^2\Big)\sigma_L^2+$$
$$\Big(K^2l^2+30L^2K^4l^2\mu_L^2\Big)\sigma_G^2 + \Big(K^2l+30L^2K^3l\mu_L^2\Big)\sum_{j=0}^{l-1}\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\|\bigg)$$

Rearranging and simplifying the terms,

$$\mathbb{E}\Big[F^c(\bar{\mathbf{x}}_C^{t+l})\Big] \leq F^c(\bar{\mathbf{x}}_C^t) - \frac{\mu_L Kl}{2}\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| + \mu_L\left(\frac{5K^2\mu_L^2 lL^2}{2}(\sigma_L^2+6K\sigma_G^2)\right)+$$
$$2\Big(L\mu_L^2+\mu_L^3KL^2(l-1)\Big)\bigg(\Big(K^2l^2+5L^2K^3l^2\mu_L^2\Big)\sigma_L^2 + \Big(K^2l^2+30L^2K^4l^2\mu_L^2\Big)\sigma_G^2\bigg)+$$
$$\left(2\Big(L\mu_L^2+\mu_L^3KL^2(l-1)\Big)\Big(K^2l+30L^2K^3l\mu_L^2\Big)+15K^3\mu_L^3 lL^2\right)\sum_{j=0}^{l-1}\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\|$$

By iterating over t, note that,

$$\sum_{t\in[T]}\mathbb{E}\Big[\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\|\Big] \leq \frac{2}{\mu_L Kl}\Big(F^c(\bar{\mathbf{x}}_C^0)-F^c(\bar{\mathbf{x}}_C^*)\Big) + \frac{T}{l}\Big(5K\mu_L^2L^2(\sigma_L^2+6K\sigma_G^2)\Big)+$$
$$\frac{4T}{l}\Big(L\mu_L+\mu_L^2KL^2(l-1)\Big)\bigg(\Big(Kl+5L^2K^2l\mu_L^2\Big)\sigma_L^2+\Big(Kl+30L^2K^3l\mu_L^2\Big)\sigma_G^2\bigg)+$$
$$\underbrace{\left(4\Big(L\mu_L+\mu_L^2KL^2(l-1)\Big)\Big(K+30L^2K^2\mu_L^2\Big)+30K^2\mu_L^2L^2\right)}_{\Gamma}\underbrace{\sum_t\sum_{j=0}^{l-1}\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\|}_{y_5} \tag{A16}$$

We bound the term y_5 as follows. We start with bounding $\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\|$ for a particular t and j:

$$\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\| = \mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\| - \mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j-1})^2\| + \mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j-1})^2\| \tag{A17}$$
$$\overset{(a_{17})}{\leq} 2\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j}) - \nabla F^c(\bar{\mathbf{x}}_C^{t+j-1})^2\| + 2\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j-1})^2\|$$
$$\overset{(a_{18})}{\leq} 2L^2\mathbb{E}\|\bar{\mathbf{x}}_C^{t+j} - \bar{\mathbf{x}}_C^{t+j-1^2}\| + 2\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j-1})^2\|$$
$$\overset{(a_{19})}{\leq} \underbrace{8L^2\mu_L^2\bigg(\Big(K^2+5L^2K^3\mu_L^2\Big)\sigma_L^2+\Big(K^2+30L^2K^4\mu_L^2\Big)\sigma_G^2\bigg)}_{B}+$$
$$\underbrace{\Big(8L^2\mu_L^2(K^2+30L^2K^3\mu_L^2)+2\Big)}_{A}\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j-1})^2\|$$

(A18)

The inequality (a_{17}) is written as a consequence of having $\mathbb{E}\|a^2\| - \mathbb{E}\|b^2\| \leq 2\mathbb{E}\|a - b^2\| + \mathbb{E}\|b^2\|$ for any random variables a and b where $a = \nabla F^c(\bar{\mathbf{x}}_C^{t+j})$ and $b = \nabla F^c(\bar{\mathbf{x}}_C^{t+j-1})$ in the inequality. The term (a_{18}) is written based on Assumption 1. Due to A3, A15, and that $l = 1$ in this case, inequality (a_{19}) was yielded. Thus:

$$\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\| \leq B + A\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j-1})^2\|$$

Unrolling recursion on j, we achieve the following:

$$\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\| \leq B\left(\frac{A^j - 1}{A - 1}\right) + A^j\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t})^2\|$$

Iterating over j, we have:

$$\sum_{j=0}^{l-1}\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\| \leq B\underbrace{\sum_{j=0}^{l-1}\left(\frac{A^j - 1}{A - 1}\right)}_{\lambda_1} + \underbrace{\left(\frac{A^l - 1}{A - 1}\right)}_{\lambda_2}\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t})^2\| \tag{A19}$$

Substituting A19 into A16:

$$\begin{aligned}&\sum_{t\in[T]}\mathbb{E}\Big[\|\nabla F^c(\bar{\mathbf{x}}_C^{t})^2\|\Big] \leq \frac{2}{\mu_L K l}\Big(F^c(\bar{\mathbf{x}}_C^{0}) - F^c(\bar{\mathbf{x}}_C^{*})\Big) + \frac{T}{l}\Big(5K\mu_L^2 L^2(\sigma_L^2 + 6K\sigma_G^2)\Big) + \\ &\frac{4T}{l}\Big(L\mu_L + \mu_L^2 K L^2(l-1)\Big)\Big(\Big(Kl + 5L^2K^2 l\mu_L^2\Big)\sigma_L^2 + \Big(Kl + 30L^2K^3 l\mu_L^2\Big)\sigma_G^2\Big) + \\ &\frac{T}{l}\Gamma\lambda_1 + \Gamma\lambda_2\sum_t \mathbb{E}\Big[\|\nabla F^c(\bar{\mathbf{x}}_C^{t})^2\|\Big]\end{aligned} \tag{A20}$$

Choosing a proper $\mu_L \leq \frac{1}{lLK2^{1.15l+1.85}}$, we have $\Gamma\lambda_2 < 1$. Thus:

$$\begin{aligned}&\min_{t\in[T]}\mathbb{E}\Big[\|\nabla F^c(\bar{\mathbf{x}}_C^{t})^2\|\Big] \leq \frac{2\big(F^c(\bar{\mathbf{x}}_C^{0}) - F^c(\bar{\mathbf{x}}_C^{*})\big)}{(1-\Gamma\lambda_2)\mu_L K T} + \frac{5K\mu_L^2 L^2(\sigma_L^2 + 6K\sigma_G^2))}{1 - \Gamma\lambda_2} + \\ &\frac{4L\mu_L + 4\mu_L^2 K L^2(l-1)}{1 - \Gamma\lambda_2}\Big(\Big(Kl + 5L^2K^2 l\mu_L^2\Big)\sigma_L^2 + \Big(Kl + 30L^2K^3 l\mu_L^2\Big)\sigma_G^2\Big) + \frac{\Gamma\lambda_1}{1 - \Gamma\lambda_2}\end{aligned} \tag{A21}$$

□

Theorem A2. *Under Assumptions 1, 2, 3, and partial participation of clients due to strategy one, if* $\mu_L \leq \frac{1}{lLK2^{1.15l+1.85}}$*, and* $t \equiv 0 \mod l$ *in Algorithm 1, the convergence rate of client model of Algorithm 1 satisfies:*

$$\begin{aligned}&\min_{t\in[T]}\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t})^2\| \leq \frac{2\big(F^c(\bar{\mathbf{x}}_C^{0}) - F^c(\bar{\mathbf{x}}_C^{*})\big)}{\mu_L K T} + \\ &\Big(5K\mu_L^2 L^2 + 4\mu_L^2 L^2\Big(K^2 l^2 + 5L^2K^3 l^2\mu_L^2\Big) + L\mu_L\Big(\frac{1}{n} + 15K^2L^2\mu_L^2\Big)\Big)\sigma_L^2 + \\ &\Big(30K^2\mu_L^2 L^2 + L\mu_L\Big(90K^3L^2\mu_L^2 + 3K\Big) + 4\mu_L^2 L^2\Big(K^2 l^2 + 30L^2K^4 l^2\mu_L^2\Big)\Big)\sigma_G^2 + \\ &\frac{\Gamma'\lambda_1}{1 - \Gamma'\lambda_2}\end{aligned}$$

Where

$$\Gamma' = 4\mu_L^2 L^2\left(K^2 l + 30L^2K^3 l\mu_L^2\right) + \frac{L\mu_L}{l}\left(90lK^3L^2\mu_L^2 + 3K\right) + 30K^2\mu_L^2L^2$$

and,

$$\bar{\mathbf{x}}_C^* = \operatorname*{argmin}_{\bar{\mathbf{x}}_C^t, t\in[T]} \mathbb{E}\left[\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\|\right]$$

Corollary A2. *Let* $\mu_L \leq \frac{1}{lLK2^{1.15l+1.85}\sqrt{T}}$. *Then, the convergence rate of the client-side model in Algorithm 1 is*

$$\min_{t\in[T]} \mathbb{E}\left[\|\nabla f^c(\bar{\mathbf{x}}_C^t)^2\|\right] \leq \mathcal{O}\left(\frac{l}{\sqrt{T}} + \frac{1}{T\sqrt{T}}\right). \tag{A22}$$

Proof. In this proof, all the gradients are w.r.t. $\mathbf{x}_C$. Due to Assumption 1, for any $\bar{\mathbf{x}}_C^{t+l}$ and $\bar{\mathbf{x}}_C^t$ such that $t \in [T]$, we can write:

$$F^c(\bar{\mathbf{x}}_C^{t+l}) \leq F^c(\bar{\mathbf{x}}_C^t) + \nabla F^c(\bar{\mathbf{x}}_C^t)^\top(\bar{\mathbf{x}}_C^{t+l} - \bar{\mathbf{x}}_C^t) + \frac{L}{2}\|\bar{\mathbf{x}}_C^{t+l} - \bar{\mathbf{x}}_C^{t}{}^2\| \tag{A23}$$

Particularly, we consider the case when $t \equiv 0 \mod l$ from now on.

Also, note the global aggregation and client update rule in the Algorithm 1 with partial worker participation,

$$\bar{\mathbf{x}}_C^{t+l} = \frac{1}{n}\sum_{i\in[S_{t+l}]} \mathbf{x}_{C,i}^{t+l} = \frac{1}{n}\sum_{i\in[S_t]} \mathbf{x}_{C,i}^{t} - \frac{\mu_L}{n}\sum_{j=0}^{l-1}\sum_{i\in[S_{t+j}]}\sum_{k=0}^{K-1}\tilde{\nabla}F_i^c(\mathbf{x}_{C,i}^{t+j,k})) =$$
$$\bar{\mathbf{x}}_C^{t} - \frac{\mu_L}{n}\sum_{j=0}^{l-1}\sum_{i\in[S_{t+j}]}\sum_{k=0}^{K-1}\tilde{\nabla}F_i^c(\mathbf{x}_{C,i}^{t+j,k}) \tag{A24}$$

Thus,

$$\bar{\mathbf{x}}_C^{t+l} - \bar{\mathbf{x}}_C^{t} = -\frac{\mu_L}{n}\sum_{j=0}^{l-1}\sum_{i\in[S_{t+j}]}\sum_{k=0}^{K-1}\tilde{\nabla}F_i^c(\mathbf{x}_{C,i}^{t+j,k}) \tag{A25}$$

In the case of partial worker participation, there are two sources of randomness. One stems from the stochastic gradient computation, while the other arises from randomly sampling the clients at round t.

Taking expectation of $F^c(\mathbf{x}_C^{t+1})$ w.r.t. both types of randomness at round $t + l - 1$, and plugging A24 into A23 note that:

$$
\begin{aligned}
&\mathbb{E}\Big[F^c(\bar{\mathbf{x}}_C^{t+l})\Big] \le F^c(\bar{\mathbf{x}}_C^t) + \Big\langle \nabla F^c(\bar{\mathbf{x}}_C^t), \mathbb{E}\Big[\bar{\mathbf{x}}_C^{t+l} - \bar{\mathbf{x}}_C^t\Big]\Big\rangle + \frac{L}{2}\mathbb{E}\|\bar{\mathbf{x}}_C^{t+l} - \bar{\mathbf{x}}_C^{t}{}^2\| \\
&\mathbb{E}\Big[F^c(\bar{\mathbf{x}}_C^{t+l})\Big] \le F^c(\bar{\mathbf{x}}_C^t) + \mu_L \underbrace{\Bigg\langle \nabla F^c(\bar{\mathbf{x}}_C^t), \mathbb{E}\Bigg[\frac{-1}{n}\sum_{j=0}^{l-1}\sum_{i\in[S_{t+j}]}\sum_{k=0}^{K-1}(\tilde{\nabla} F_i^c(\mathbf{x}_{C,i}^{t+j,k}) - \nabla F_i^c(\bar{\mathbf{x}}_C^{t+j}))\Bigg]\Bigg\rangle}_{A_1'} \\
&\underbrace{-\mu_L\Bigg\langle \nabla F^c(\bar{\mathbf{x}}_C^t), \mathbb{E}\Bigg[\frac{K}{n}\sum_{j=0}^{l-1}\sum_{i\in[S_{t+j}]}\nabla F_i^c(\bar{\mathbf{x}}_C^{t+j})\Bigg]\Bigg\rangle}_{A_2'} + \frac{L\mu_L^2}{2}\underbrace{\mathbb{E}\Bigg[\|\frac{1}{n}\sum_{j=0}^{l-1}\sum_{i\in[S_{t+j}\|]}\sum_{k=0}^{K-1}\tilde{\nabla} F_i^c(\mathbf{x}_{C,i}^{t+j,k})^2\Bigg]}_{A_3'}
\end{aligned}
\tag{A26}
$$

According to [15, Lemma 1], terms A_1' and A_2' will possess the same bounds as those of A_1 and A_2. Thus:

$$
A_1' \le \frac{K}{2}\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| + \frac{5K^2\mu_L^2 lL^2}{2}(\sigma_L^2 + 6K\sigma_G^2) + 15K^3\mu_L^2 lL^2\sum_{j=0}^{l-1}\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\| \tag{A27}
$$

$$
\begin{aligned}
A_2' \le &-\frac{\mu_L K(l+1)}{2}\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| + \frac{\mu_L^3 KL^2(l-1)}{2}\Bigg(4\Big(K^2l^2 + 5L^2K^3l^2\mu_L^2\Big)\sigma_L^2 \\
&+ 4\Big(K^2l^2 + 30L^2K^4l^2\mu_L^2\Big)\sigma_G^2 + 4\Big(K^2l + 30L^2K^3l\mu_L^2\Big)\sum_{j=0}^{l-1}\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\|\Bigg)
\end{aligned}
\tag{A28}
$$

Note that $[S_t] = \{q_1^t, ..., q_n^t\}$ is the index set demonstrating the sampled clients, which might contain duplicate elements, as the sampling is with replacement. We now proceed to bound the term A_3' following [15]:

$$
\begin{aligned}
A_3' &= \mathbb{E}\Bigg[\|\frac{1}{n}\sum_{j=0}^{l-1}\sum_{i\in[S_{t+j}]}\sum_{k=0}^{K-1}\tilde{\nabla} F_i^c(\mathbf{x}_{C,i}^{t+j,k})^2\|\Bigg] \\
&= \frac{1}{n^2}\mathbb{E}\Bigg[\|\sum_{j=0}^{l-1}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\tilde{\nabla} F_{q_i^{t+j}}^c(\mathbf{x}_{C,q_i^{t+j}}^{t+j,k})^2\|\Bigg] \\
&\overset{a_1'}{=} \frac{1}{n^2}\mathbb{E}\Bigg[\|\sum_{j=0}^{l-1}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\Big(\tilde{\nabla} F_{q_i^{t+j}}^c(\mathbf{x}_{C,q_i^{t+j}}^{t+j,k}) - \nabla F_{q_i^{t+j}}^c(\mathbf{x}_{C,q_i^{t+j}}^{t+j,k})\Big)^2\|\Bigg] \\
&+ \frac{1}{n^2}\mathbb{E}\Bigg[\|\sum_{j=0}^{l-1}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\nabla F_{q_i^{t+j}}^c(\mathbf{x}_{C,q_i^{t+j}}^{t+j,k})^2\|\Bigg] \\
&\overset{a_2'}{\le} \frac{lK\sigma_L^2}{n} + \frac{1}{n^2}\mathbb{E}\Bigg[\|\sum_{j=0}^{l-1}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\nabla F_{q_i^{t+j}}^c(\mathbf{x}_{C,q_i^{t+j}}^{t+j,k})^2\|\Bigg]
\end{aligned}
\tag{A29}
$$

The equality a_1' follows from the fact that $\mathbb{E}\left[\|z^2\|\right] = \mathbb{E}\left[\|z - \mathbb{E}[z]^2\|\right] + \|\mathbb{E}\left[z\right]^2\|$ and the inequality a_2' is due to assumption 3 and the explanation provided in A10. Now, let's consider $\mathbf{t}_i^j = \sum_{k=0}^{K-1} \nabla F_i^c(\mathbf{x}_{C,i}^{t+j,k})$, then:

$$
\begin{aligned}
&\mathbb{E}\left[\|\sum_{j=0}^{l-1}\sum_{i=1}^{n}\sum_{k=0}^{K-1} \nabla F^c_{q_i^{t+j}}(\mathbf{x}^{t+j,k}_{C,q_i^{t+j}})^2\|\right] = \mathbb{E}\left[\|\sum_{j=0}^{l-1}\sum_{i=1}^{n} \mathbf{t}^j_{q_i^{t+j}}{}^2\|\right] \\
&= \mathbb{E}\left[\sum_{j=0}^{l-1}\sum_{i=1}^{n} \|\mathbf{t}^j_{q_i^{t+j}}{}^2\| + \sum_{j=0}^{l-1} \sum_{i \neq z, q_i^{t+j}, q_z^{t+j} \in [S_{t+j}]} \langle \mathbf{t}^j_{q_i^{t+j}}, \mathbf{t}^j_{q_z^{t+j}} \rangle\right] \\
&\overset{a_3'}{=} \mathbb{E}\left[n \sum_{j=0}^{l-1} \|\mathbf{t}^j_{q_1^{t+j}}{}^2\| + n(n-1) \sum_{j=0}^{l-1} \langle \mathbf{t}^j_{q_1^{t+j}}, \mathbf{t}^j_{q_2^{t+j}} \rangle\right] \\
&= \frac{n}{m} \sum_{j=0}^{l-1} \sum_{i \in [S]} \|\mathbf{t}_i^{j2}\| + \frac{n(n-1)}{m^2} \sum_{j=0}^{l-1} \sum_{i,z \in [S]} \langle \mathbf{t}_i^j, \mathbf{t}_z^j \rangle \\
&= \frac{n}{m} \sum_{j=0}^{l-1} \sum_{i \in [S]} \|\mathbf{t}_i^{j2}\| + \frac{n(n-1)}{m^2} \sum_{j=0}^{l-1} \|\sum_{i \in [S]} \mathbf{t}_i^{j2}\| \\
&\overset{a_4'}{\leq} \frac{n^2}{m} \underbrace{\sum_{j=0}^{l-1} \sum_{i \in [S]} \|\mathbf{t}_i^{j2}\|}_{A_4'}
\end{aligned}
\tag{A30}
$$

Note that the equality a_3' is due to independent sampling with replacement as outlined by [15] and a_4' follows from A8. Now, we bound the term A_4' as follows:

$$
\begin{aligned}
&\sum_{j=0}^{l-1} \sum_{i \in [S]} \|\mathbf{t}_i^{j2}\| = \sum_{j=0}^{l-1} \sum_{i \in [S]} \|\sum_{k=0}^{K-1} \nabla F_i^c(\mathbf{x}_{C,i}^{t+j,k})^2\| \overset{a_5'}{=} K \sum_{j=0}^{l-1} \sum_{i \in [S]} \sum_{k=0}^{K-1} \|\nabla F_i^c(\mathbf{x}_{C,i}^{t+j,k})^2\| \\
&= K \sum_{j=0}^{l-1} \sum_{i \in [S]} \sum_{k=0}^{K-1} \|\nabla F_i^c(\mathbf{x}_{C,i}^{t+j,k}) - \nabla F_i^c(\mathbf{x}_C^{t+j}) + \nabla F_i^c(\mathbf{x}_C^{t+j}) - \nabla F^c(\mathbf{x}_C^{t+j}) + \nabla F^c(\mathbf{x}_C^{t+j})^2\| \\
&\overset{a_6'}{\leq} 3KL^2 \sum_{j=0}^{l-1} \sum_{i \in [S]} \sum_{k=0}^{K-1} \|\mathbf{x}_{C,i}^{t+j,k} - \mathbf{x}_C^{t+j2}\| + 3mlK^2\sigma_G^2 + 3mK^2 \sum_{j=0}^{l-1} \|\nabla F^c(\mathbf{x}_C^{t+j})^2\| \\
&\overset{a_7'}{\leq} 15mlK^3L^2\mu_L^2(\sigma_L^2 + 6K\sigma_G^2) + 3mlK^2\sigma_G^2 \\
&+ (90mlK^4L^2\mu_L^2 + 3mK^2) \sum_{j=0}^{l-1} \mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\|
\end{aligned}
\tag{A31}
$$

The term a_5' follows from the fact A8, the term a_6' stems from the fact A8 and the assumption 1. The term a_7' is due to A11. Now, plugging A31 into A30 and A30 into A29, we have the following bound on A_3':

$$
\begin{aligned}
&\mathbb{E}\left[\|\frac{1}{n}\sum_{j=0}^{l-1} \sum_{i \in [S_{t+j}]} \sum_{k=0}^{K-1} \tilde{\nabla} F_i^c(\mathbf{x}_{C,i}^{t+j,k})^2\|\right] \\
&\leq \left(\frac{lK}{n} + 15lK^3L^2\mu_L^2\right)\sigma_L^2 + \left(90lK^4L^2\mu_L^2 + 3lK^2\right)\sigma_G^2 \\
&+ (90lK^4L^2\mu_L^2 + 3K^2) \sum_{j=0}^{l-1} \mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\|
\end{aligned}
\tag{A32}
$$

doi:10.20944/preprints202409.0335.v1

Plugging A27, A28 and A32 into A23, with rearrangement and simplification, observe that:

$$\begin{aligned}
&\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| \leq \frac{2}{\mu_L K l}\Big(-\mathbb{E}\Big[F^c(\bar{\mathbf{x}}_C^{t+l})\Big] + F^c(\bar{\mathbf{x}}_C^t)\Big) + \\
&\Big(5K\mu_L^2L^2 + 4\mu_L^2L^2\Big(K^2l^2 + 5L^2K^3l^2\mu_L^2\Big) + L\mu_L\Big(\frac{1}{n} + 15K^2L^2\mu_L^2\Big)\Big)\sigma_L^2 + \\
&\Big(30K^2\mu_L^2L^2 + L\mu_L\Big(90K^3L^2\mu_L^2 + 3K\Big) + 4\mu_L^2L^2\Big(K^2l^2 + 30L^2K^4l^2\mu_L^2\Big)\Big)\sigma_G^2 + \\
&\underbrace{\Big(4\mu_L^2L^2\Big(K^2l + 30L^2K^3l\mu_L^2\Big) + \frac{L\mu_L}{l}\Big(90lK^3L^2\mu_L^2 + 3K\Big) + 30K^2\mu_L^2L^2\Big)}_{\Gamma'}\sum_{j=0}^{l-1}\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\| \qquad \text{(A33)}
\end{aligned}$$

By iterating over t, note that,

$$\begin{aligned}
&\sum_{t\in[T]}\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| \leq \frac{2}{\mu_L K l}\Big(-\mathbb{E}\Big[F^c(\bar{\mathbf{x}}_C^*)\Big] + F^c(\bar{\mathbf{x}}_C^0)\Big) + \\
&\frac{T}{l}\Big(5K\mu_L^2L^2 + 4\mu_L^2L^2\Big(K^2l^2 + 5L^2K^3l^2\mu_L^2\Big) + L\mu_L\Big(\frac{1}{n} + 15K^2L^2\mu_L^2\Big)\Big)\sigma_L^2 + \\
&\frac{T}{l}\Big(30K^2\mu_L^2L^2 + L\mu_L\Big(90K^3L^2\mu_L^2 + 3K\Big) + 4\mu_L^2L^2\Big(K^2l^2 + 30L^2K^4l^2\mu_L^2\Big)\Big)\sigma_G^2 + \\
&\Gamma'\sum_t\sum_{j=0}^{l-1}\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^{t+j})^2\| \qquad \text{(A34)}
\end{aligned}$$

Due to A19, observe that:

$$\begin{aligned}
&\sum_{t\in[T]}\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| \leq \frac{2}{\mu_L K l}\Big(-\mathbb{E}\Big[F^c(\bar{\mathbf{x}}_C^*)\Big] + F^c(\bar{\mathbf{x}}_C^0)\Big) + \\
&\frac{T}{l}\Big(5K\mu_L^2L^2 + 4\mu_L^2L^2\Big(K^2l^2 + 5L^2K^3l^2\mu_L^2\Big) + L\mu_L\Big(\frac{1}{n} + 15K^2L^2\mu_L^2\Big)\Big)\sigma_L^2 + \\
&\frac{T}{l}\Big(30K^2\mu_L^2L^2 + L\mu_L\Big(90K^3L^2\mu_L^2 + 3K\Big) + 4\mu_L^2L^2\Big(K^2l^2 + 30L^2K^4l^2\mu_L^2\Big)\Big)\sigma_G^2 + \\
&\frac{T}{l}\Gamma'\lambda_1 + \Gamma'\lambda_2\sum_t\mathbb{E}\Big[\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\|\Big] \qquad \text{(A35)}
\end{aligned}$$

We let $\mu_L \leq \frac{1}{lLK2^{1.15l+1.85}}$, thus:

$$\begin{aligned}
&\min_{t\in[T]}\mathbb{E}\|\nabla F^c(\bar{\mathbf{x}}_C^t)^2\| \leq \frac{2\big(F^c(\bar{\mathbf{x}}_C^0) - F^c(\bar{\mathbf{x}}_C^*)\big)}{\mu_L K T} + \\
&\Big(5K\mu_L^2L^2 + 4\mu_L^2L^2\Big(K^2l^2 + 5L^2K^3l^2\mu_L^2\Big) + L\mu_L\Big(\frac{1}{n} + 15K^2L^2\mu_L^2\Big)\Big)\sigma_L^2 + \\
&\Big(30K^2\mu_L^2L^2 + L\mu_L\Big(90K^3L^2\mu_L^2 + 3K\Big) + 4\mu_L^2L^2\Big(K^2l^2 + 30L^2K^4l^2\mu_L^2\Big)\Big)\sigma_G^2 + \\
&\frac{\Gamma'\lambda_1}{1 - \Gamma'\lambda_2} \qquad \text{(A36)}
\end{aligned}$$

This completes the proof.

□

Appendix A.2. Server-Side Model Convergence

Theorem A3. *Under Assumptions 1, 2, 3, and full participation of clients, if $\mu \leq \frac{1}{8Lm^2}$, $t \equiv 0 \mod l$, the convergence rate of the server model of Algorithm 1 satisfies:*

$$\min_t \mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| \leq \frac{2l\left(F^s(\mathbf{x}_S^0) - F^s(\mathbf{x}_S^*)\right)}{\mu(2m-3)T} + \frac{L\mu m^2}{2m-3}\left(9.2\sigma_L^2 + 13.2\sigma_G^2\right)$$

Corollary A3. *Let $\mu \leq \frac{1}{Lm^2\sqrt{T}}$, then the convergence rate of the server-side model is:*

$$\min_{t\in[T]} \mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| \leq \mathcal{O}\left(\frac{l}{\sqrt{T}}\right)$$

Proof. Due to Assumption 1, for any $\mathbf{x}_S^{t+l}$ and $\mathbf{x}_S^t$, it can be written that:

$$F^s(\mathbf{x}_S^{t+l}) \leq F^s(\mathbf{x}_S^t) + \nabla F^s(\mathbf{x}_S^t)^\top(\mathbf{x}_S^{t+l} - \mathbf{x}_S^t) + \frac{L}{2}\|\mathbf{x}_S^{t+l} - \mathbf{x}_S^t{}^2\| \tag{A37}$$

Also, note the client forward-propagation and server model update rules in the Algorithm 1,

$$\mathbf{x}_{S,i+1}^t = \mathbf{x}_{S,i}^t - \mu\tilde{\nabla}F_i^s(\mathbf{x}_{S,i}^t) \tag{A38}$$

Thus, putting $\mathbf{x}_S^t = \mathbf{x}_{S,0}^t$ and $\mathbf{x}_S^{t+l} = \mathbf{x}_{S,m}^t$, note that:

$$\mathbf{x}_S^{t+l} - \mathbf{x}_S^t = -\mu\sum_{i=0}^{m-1}\tilde{\nabla}F_i^s(\mathbf{x}_{S,i}^t) \tag{A39}$$

Taking expectation with respect to randomness at round t, i.e., ¸$^{[t]} \triangleq [\zeta_i^\tau]_{i\in[N],\tau\in[t]}$, and plugging A38 into A37 note that:

$$\begin{aligned}
\mathbb{E}\left[F^s(\mathbf{x}_S^{t+l})\right] &\leq F^s(\mathbf{x}_S^t) - \mu\left\langle\nabla F^s(\mathbf{x}_S^t), \mathbb{E}\left[\sum_{i=0}^{m-1}\tilde{\nabla}F_i^s(\mathbf{x}_{S,i}^t)\right]\right\rangle + \frac{L\mu^2}{2}\mathbb{E}\|\sum_{i=0}^{m-1}\tilde{\nabla}F_i^s(\mathbf{x}_{S,i}^t)^2\| \\
\mathbb{E}\left[F^s(\mathbf{x}_S^{t+l})\right] &\leq F^s(\mathbf{x}_S^t) \underbrace{-\mu\left\langle\nabla F^s(\mathbf{x}_S^t), \mathbb{E}\left[\sum_{i=0}^{m-1}\left(\tilde{\nabla}F_i^s(\mathbf{x}_{S,i}^t) - \nabla F_i^s(\mathbf{x}_S^t)\right)\right]\right\rangle}_{B_1} \\
&\underbrace{-\mu\left\langle\nabla F^s(\mathbf{x}_S^t), \mathbb{E}\left[\sum_{i=0}^{m-1}\nabla F_i^s(\mathbf{x}_S^t)\right]\right\rangle}_{B_2} + \frac{L\mu^2}{2}\underbrace{\mathbb{E}\|\sum_{i=0}^{m-1}\tilde{\nabla}F_i^s(\mathbf{x}_{S,i}^t)^2\|}_{B_3}
\end{aligned} \tag{A40}$$

The term B_1 will be bounded as follows:

$$\begin{aligned}
&-\mu\left\langle \nabla F^s(\mathbf{x}_S^t), \mathbb{E}\left[\mathbb{E}\left[\sum_{i=0}^{m-1}\left(\tilde{\nabla} F_i^s(\mathbf{x}_{S,i}^t) - \nabla F_i^s(\mathbf{x}_S^t)\right)\middle|\xi\right]\right]\right\rangle \overset{(b_1)}{=} \frac{\mu}{2}\mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\|+ \\
&\frac{\mu}{2}\mathbb{E}\|\sum_{i=0}^{m-1}\left(\nabla F_i^s(\mathbf{x}_{S,i}^t) - \nabla F_i^s(\mathbf{x}_S^t)\right)^2\| - \frac{\mu}{2}\mathbb{E}\|\sum_{i=0}^{m-1}\left(\nabla F_i^s(\mathbf{x}_{S,i}^t) - \nabla F_i^s(\mathbf{x}_S^t)\right) + \nabla F^s(\mathbf{x}_S^t)^2\| \\
&\overset{(b_2)}{\leq} \frac{\mu}{2}\mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| + \frac{\mu}{2}\mathbb{E}\|\sum_{i=0}^{m-1}\left(\nabla F_i^s(\mathbf{x}_{S,i}^t) - \nabla F_i^s(\mathbf{x}_S^t)\right)^2\| \\
&\overset{(b_3)}{\leq} \frac{\mu}{2}\mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| + \frac{m\mu}{2}\sum_{i=0}^{m-1}\mathbb{E}\|\left(\nabla F_i^s(\mathbf{x}_{S,i}^t) - \nabla F_i^s(\mathbf{x}_S^t)\right)^2\| \\
&\overset{(b_4)}{\leq} \frac{\mu}{2}\mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| + \frac{m\mu L^2}{2}\sum_{i=0}^{m-1}\mathbb{E}\|\mathbf{x}_{S,i}^t - \mathbf{x}_S^{t\,2}\| \qquad \text{(A41)}
\end{aligned}$$

The equality (b_1) is due to $-\langle a, b\rangle = \frac{1}{2}\left(\|a^2\| + \|b^2\| - \|a+b^2\|\right)$ for any two vectors a and b. The inequality (b_2) is clear as we dropped a negative term, inequality (b_3) stems from the fact that $\mathbb{E}\|\sum_i^n z_i{}^2\| \leq n\sum_i \mathbb{E}\left[\|z_i{}^2\|\right]$ holds for any random variable z_i, and the inequality (b_4) is due to 1.

The term B_2 will be bounded as follows:

$$-\mu\left\langle \nabla F^s(\mathbf{x}_S^t), \mathbb{E}\left[\sum_{i=0}^{m-1}\nabla F_i^s(\mathbf{x}_S^t)\right]\right\rangle \overset{(b_5)}{=} -m\mu\mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| \qquad \text{(A42)}$$

Note that the equality (b_5) holds based on the definition 3.

The term B_3 will be bounded as below:

$$\begin{aligned}
&\mathbb{E}\|\sum_{i=0}^{m-1}\tilde{\nabla} F_i^s(\mathbf{x}_{S,i}^t)^2\| \overset{(b_6)}{\leq} m\sum_{i=0}^{m-1}\mathbb{E}\|\tilde{\nabla} F_i^s(\mathbf{x}_{S,i}^t)^2\| \\
&= m\sum_{i=0}^{m-1}\mathbb{E}\|\tilde{\nabla} F_i^s(\mathbf{x}_{S,i}^t) - \nabla F_i^s(\mathbf{x}_{S,i}^t) + \nabla F_i^s(\mathbf{x}_{S,i}^t)^2\| \\
&\overset{(b_7)}{\leq} 2m\sum_{i=0}^{m-1}\mathbb{E}\|\tilde{\nabla} F_i^s(\mathbf{x}_{S,i}^t) - \nabla F_i^s(\mathbf{x}_{S,i}^t)^2\| + 2m\sum_{i=0}^{m-1}\mathbb{E}\|\nabla F_i^s(\mathbf{x}_{S,i}^t)^2\| \\
&\overset{(b_8)}{\leq} 2m^2\sigma_L^2 + 2m\sum_{i=0}^{m-1}\mathbb{E}\|\nabla F_i^s(\mathbf{x}_{S,i}^t) - \nabla F^s(\mathbf{x}_{S,i}^t) + \nabla F^s(\mathbf{x}_{S,i}^t) - \nabla F^s(\mathbf{x}_S^t) + \nabla F^s(\mathbf{x}_S^t)^2\| \\
&\overset{(b_9)}{\leq} 2m^2\sigma_L^2 + 6m\sum_{i=0}^{m-1}\mathbb{E}\|\nabla F_i^s(\mathbf{x}_{S,i}^t) - \nabla F^s(\mathbf{x}_{S,i}^t)^2\| + 6m\sum_{i=0}^{m-1}\mathbb{E}\|\nabla F^s(\mathbf{x}_{S,i}^t) - \nabla F^s(\mathbf{x}_S^t)^2\|+ \\
&6m\sum_{i=0}^{m-1}\mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| \\
&\overset{(b_{10})}{\leq} 2m^2\sigma_L^2 + 6m^2\sigma_G^2 + 6mL^2\underbrace{\sum_{i=0}^{m-1}\mathbb{E}\|\mathbf{x}_{S,i}^t - \mathbf{x}_S^{t\,2}\|}_{B_4} + 6m^2\mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| \qquad \text{(A43)}
\end{aligned}$$

The inequalities (b_6), (b_7), and (b_9) due to the same reason (b_3) holds above. The inequalities (b_8) and (b_{10}) hold due to Assumptions 3. The term B_4 is bounded similar to [13, Lemma 3]:

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{x}^t_{S,i} - \mathbf{x}^t_S{}^2\| \overset{(b_{11})}{\leq} \mathbb{E}\|\mathbf{x}^t_{S,i-1}\| - \mathbf{x}^t_S - \mu\tilde{\nabla}F^s_{i-1}(\mathbf{x}^t_{S,i-1})^2 \\
&= \mathbb{E}\|\mathbf{x}^t_{S,i-1}\| - \mathbf{x}^t_S{}^2 + \mathbb{E}\|\mu\tilde{\nabla}F^s_{i-1}(\mathbf{x}^t_{S,i-1})^2\| + 2\Big\langle \mathbf{x}^t_{S,i-1} - \mathbf{x}^t_S, -\mu\tilde{\nabla}F^s_{i-1}(\mathbf{x}^t_{S,i-1})\Big\rangle \\
&\overset{(b_{12})}{\leq} (1 + \frac{1}{2m-1})\mathbb{E}\|\mathbf{x}^t_{S,i-1} - \mathbf{x}^t_S{}^2\| + (1+2m)\mathbb{E}\|\mu\tilde{\nabla}F^s_{i-1}(\mathbf{x}^t_{S,i-1})^2\| \\
&= (1 + \frac{1}{2m-1})\mathbb{E}\|\mathbf{x}^t_{S,i-1} - \mathbf{x}^t_S{}^2\| + (1+2m)\mu^2\mathbb{E}\Bigg\|\tilde{\nabla}F^s_{i-1}(\mathbf{x}^t_{S,i-1}) - \nabla F^s_{i-1}(\mathbf{x}^t_{S,i-1}) + \\
&\nabla F^s_{i-1}(\mathbf{x}^t_{S,i-1}) - \nabla F^s_{i-1}(\mathbf{x}^t_S) + \nabla F^s_{i-1}(\mathbf{x}^t_S) - \nabla F^s(\mathbf{x}^t_S) + \nabla F^s(\mathbf{x}^t_S)\Bigg\|^2 \\
&\overset{(b_{13})}{\leq} (1 + \frac{1}{2m-1})\mathbb{E}\|\mathbf{x}^t_{S,i-1} - \mathbf{x}^t_S{}^2\| + 4(1+2m)\mu^2\mathbb{E}\|\tilde{\nabla}F^s_{i-1}(\mathbf{x}^t_{S,i-1}) - \nabla F^s_{i-1}(\mathbf{x}^t_{S,i-1})^2\| + \\
&4(1+2m)\mu^2\mathbb{E}\|\nabla F^s_{i-1}(\mathbf{x}^t_{S,i-1}) - \nabla F^s_{i-1}(\mathbf{x}^t_S)^2\| + 4(1+2m)\mu^2\mathbb{E}\|\nabla F^s_{i-1}(\mathbf{x}^t_S\|) - \nabla F^s(\mathbf{x}^t_S)^2 \\
&+ 4(1+2m)\mu^2\mathbb{E}\|\nabla F^s(\mathbf{x}^t_S)^2\| \\
&\overset{(b_{14})}{\leq} (1 + \frac{1}{2m-1} + 4(1+2m)\mu^2L^2)\mathbb{E}\|\mathbf{x}^t_{S,i-1} - \mathbf{x}^t_S{}^2\| + 4(1+2m)\mu^2(\sigma_L^2 + \sigma_G^2) + \\
&4(1+2m)\mu^2\mathbb{E}\|\nabla F^s(\mathbf{x}^t_S)^2\|
\end{aligned}
$$

The A38 implies the inequality (b_{11}). The inequality (b_{12}) holds true based on $2\langle a, b\rangle \leq \frac{1}{n-1}\|a^2\| + n\|b^2\|$ for any two vectors a, b and positive number n. The inequality (b_{13}) follows from the previously mentioned fact at the inequality (b_3), and the inequality (b_{14}) is based on Assumptions 3 and 1. Given $\mu \leq \frac{1}{2L(2m+1)}$ and by averaging over the clients, observe that:

$$
\begin{aligned}
&\frac{1}{m}\sum_{i=0}^{m-1}\mathbb{E}\|\mathbf{x}^t_{S,i} - \mathbf{x}^t_S{}^2\| \leq (1 + \frac{4m}{4m^2-1})\frac{1}{m}\sum_{i=1}^{m-1}\mathbb{E}\|\mathbf{x}^t_{S,i-1} - \mathbf{x}^t_S{}^2\| + 4(1+2m)\mu^2(\sigma_L^2 + \sigma_G^2) + \\
&4(1+2m)\mu^2\mathbb{E}\|\nabla F^s(\mathbf{x}^t_S)^2\| \\
&\leq (1 + \frac{1}{m-1})\frac{1}{m}\sum_{i=1}^{m-1}\mathbb{E}\|\mathbf{x}^t_{S,i-1} - \mathbf{x}^t_S{}^2\| + 4(1+2m)\mu^2(\sigma_L^2 + \sigma_G^2) + 4(1+2m)\mu^2\mathbb{E}\|\nabla F^s(\mathbf{x}^t_S)^2\|
\end{aligned}
$$

Unrolling the recursion, following [13, Lemma 3], it is inferred that:

$$
\begin{aligned}
&\frac{1}{m}\sum_{i=0}^{m-1}\mathbb{E}\|\mathbf{x}^t_{S,i} - \mathbf{x}^t_S{}^2\| \\
&\leq \sum_{j=0}^{m-1}(1 + \frac{1}{m-1})^j\Big(4(1+2m)\mu^2(\sigma_L^2 + \sigma_G^2) + 4(1+2m)\mu^2\mathbb{E}\|\nabla F^s(\mathbf{x}^t_S)^2\|\Big) \\
&\leq (m-1) \times \left[\left(1 + \frac{1}{m-1}\right)^m - 1\right] \times \left[4(1+2m)\mu^2(\sigma_L^2 + \sigma_G^2) + 4(1+2m)\mu^2\mathbb{E}\|\nabla F^s(\mathbf{x}^t_S)^2\|\right] \\
&\leq 16(m+2m^2)\mu^2(\sigma_L^2 + \sigma_G^2) + 16(m+2m^2)\mu^2\mathbb{E}\|\nabla F^s(\mathbf{x}^t_S)^2\| \qquad \text{(A44)}
\end{aligned}
$$

Note that in the above inequality, $\left(1+\frac{1}{m-1}\right)^m - 1 \leq 4$ for $m > 1$.

Plugging A41, A42, A43 and A44 into A40, observe that :

$$\mathbb{E}\Big[F^s(\mathbf{x}_S^{t+l})\Big] \leq F^s(\mathbf{x}_S^t) + \frac{1}{2}\Big(16L^2\mu^3m^3(1+2m) + 2L\mu^2m^2 + 96L^3\mu^4m^3(1+2m)\Big)\sigma_L^2$$
$$+ \frac{1}{2}\Big(16L^2\mu^3m^3(1+2m) + 6L\mu^2m^2 + 96L^3\mu^4m^3(1+2m)\Big)\sigma_G^2 +$$
$$\frac{1}{2}\Big(-2\mu m + 6Lm^2\mu^2 + \mu + 96L^3\mu^4m^3(1+2m) + 16L^2\mu^3m^3(1+2m)\Big)\mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\|$$
$$\overset{(b_{15})}{\leq} F^s(\mathbf{x}_S^t) + \frac{\mu(3-2m)}{2}\mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| + \frac{L\mu^2m^2}{2}\Big(9.2\sigma_L^2 + 13.2\sigma_G^2\Big)$$

Rearranging the terms, and summing over t, observe that:

$$\sum_t \mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| \leq \frac{2\Big(F^s(\mathbf{x}_S^0) - F^s(\mathbf{x}_S^*)\Big)}{\mu(2m-3)} + \sum_t \frac{L\mu m^2}{2m-3}\Big(9.2\sigma_L^2 + 13.2\sigma_G^2\Big) \tag{A45}$$

Assuming there are T global rounds overall,

$$\min_t \mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| \leq \frac{2l\Big(F^s(\mathbf{x}_S^0) - F^s(\mathbf{x}_S^*)\Big)}{\mu(2m-3)T} + \frac{L\mu m^2}{2m-3}\Big(9.2\sigma_L^2 + 13.2\sigma_G^2\Big) \tag{A46}$$

□

Theorem A4. *Under Assumptions 1, 2, 3, and full participation of clients, if $\mu \leq \frac{1}{8L^2m^2}$, $t \equiv 0 \mod l$, the convergence rate of the server model of Algorithm 1 satisfies:*

$$\min_t \mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| \leq \frac{l\Big(F^s(\mathbf{x}_S^0) - F^s(\mathbf{x}_S^*)\Big)}{\mu(m-2)T} + \frac{L\mu m^2}{m-2}\Big(7\sigma_L^2 + 7\sigma_G^2\Big)$$

Corollary A4. *Let $\mu \leq \frac{1}{L^2m^2\sqrt{T}}$, then the convergence rate of the server-side model is:*

$$\min_{t\in[T]} \mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| \leq \mathcal{O}\Big(\frac{l}{\sqrt{T}}\Big)$$

Proof. Due to Assumption 1, for any $\mathbf{x}_S^{t+l}$ and $\mathbf{x}_S^t$, it can be written that:

$$F^s(\mathbf{x}_S^{t+l}) \leq F^s(\mathbf{x}_S^t) + \nabla F^s(\mathbf{x}_S^t)^\top(\mathbf{x}_S^{t+l} - \mathbf{x}_S^t) + \frac{L}{2}\|\mathbf{x}_S^{t+l} - \mathbf{x}_S^{t\,2}\| \tag{A47}$$

Also, note the client forward-propagation and server model update rules in the Algorithm 1,

$$\mathbf{x}_{S,i+1}^t = \mathbf{x}_{S,i}^t - \mu\tilde{\nabla}F_i^s(\mathbf{x}_{S,i}^t) \tag{A48}$$

Thus, putting $\mathbf{x}_S^t = \mathbf{x}_{S,0}^t$ and $\mathbf{x}_S^{t+l} = \mathbf{x}_{S,m}^t$, note that:

$$\mathbf{x}_S^{t+l} - \mathbf{x}_S^t = -\mu\sum_{i=0}^{m-1}\tilde{\nabla}F_i^s(\mathbf{x}_{S,i}^t) \tag{A49}$$

Taking expectation for both types of randomness, i.e., randomness due to stochastic gradients and due to sampling of clients, at round t, and plugging A48 into A47 note that:

doi:10.20944/preprints202409.0335.v1

$$
\begin{aligned}
&\mathbb{E}\Big[F^s(\mathbf{x}_S^{t+l})\Big] \leq F^s(\mathbf{x}_S^t) - \mu\Big\langle \nabla F^s(\mathbf{x}_S^t), \mathbb{E}\Big[\sum_{i\in[S_t]} \tilde{\nabla} F_i^s(\mathbf{x}_{S,i}^t)\Big]\Big\rangle + \frac{L\mu^2}{2}\mathbb{E}\|\sum_{i\in[S_t]} \tilde{\nabla} F_i^s(\mathbf{x}_{S,i}^t)^2\| \\
&\mathbb{E}\Big[F^s(\mathbf{x}_S^{t+l})\Big] \leq F^s(\mathbf{x}_S^t) \underbrace{-\mu\Bigg\langle \nabla F^s(\mathbf{x}_S^t), \mathbb{E}\Bigg[\sum_{i\in[S_t]} (\tilde{\nabla} F_i^s(\mathbf{x}_{S,i}^t) - \nabla F_i^s(\mathbf{x}_S^t))\Bigg]\Bigg\rangle}_{B_1} \\
&\underbrace{-\mu\Bigg\langle \nabla F^s(\mathbf{x}_S^t), \mathbb{E}\Bigg[\sum_{i\in[S_t]} \nabla F_i^s(\mathbf{x}_S^t)\Bigg]\Bigg\rangle}_{B_2} + \frac{L\mu^2}{2}\underbrace{\mathbb{E}\|\sum_{i\in[S_t]} \tilde{\nabla} F_i^s(\mathbf{x}_{S,i}^t)^2\|}_{B_3}
\end{aligned}
\tag{A50}
$$

Note that $[S_t] = \{q_1^t, ..., q_n^t\}$ is the index set demonstrating the sampled clients, which might contain duplicate elements, as the sampling is with replacement. The term B_1 will be bounded as follows:

$$
\begin{aligned}
&-\mu\Bigg\langle \nabla F^s(\mathbf{x}_S^t), \mathbb{E}\Bigg[\mathbb{E}\Bigg[\sum_{i\in[S_t]} (\tilde{\nabla} F_i^s(\mathbf{x}_{S,i}^t) - \nabla F_i^s(\mathbf{x}_S^t))\Bigg|\xi\Bigg]\Bigg]\Bigg\rangle \overset{(b_1)}{=} \frac{\mu}{2}\mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| + \\
&\frac{\mu}{2}\mathbb{E}\|\sum_{i\in[S_t]} (\nabla F_i^s(\mathbf{x}_{S,i}^t) - \nabla F_i^s(\mathbf{x}_S^t))^2\| - \frac{\mu}{2}\mathbb{E}\|\sum_{i\in[S_t]} (\nabla F_i^s(\mathbf{x}_{S,i}^t) - \nabla F_i^s(\mathbf{x}_S^t)) + \nabla F^s(\mathbf{x}_S^t)^2\| \\
&\overset{(b_2)}{\leq} \frac{\mu}{2}\mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| + \frac{\mu}{2}\mathbb{E}\|\sum_{i\in[S_t]} (\nabla F_i^s(\mathbf{x}_{S,i}^t) - \nabla F_i^s(\mathbf{x}_S^t))^2\| \\
&\overset{(b_3)}{\leq} \frac{\mu}{2}\mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| + \frac{\mu}{2}\mathbb{E}\|\sum_{i=1}^{n} (\nabla F_{q_i^t}^s(\mathbf{x}_{S,q_i^t}^t) - \nabla F_{q_i^t}^s(\mathbf{x}_S^t))^2\| \\
&\overset{(b_4)}{\leq} \frac{\mu}{2}\mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| + \frac{n\mu L^2}{2}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{x}_{S,q_i^t}^t - \mathbf{x}_S^{t\,2}\|
\end{aligned}
\tag{A51}
$$

The equality (b_1) is due to $-\langle a, b\rangle = \frac{1}{2}(\|a^2\| + \|b^2\| - \|a + b^2\|)$ for any two vectors a and b. The inequality (b_2) is clear as we dropped a negative term, inequality (b_4) stems from the fact that $\mathbb{E}\|\sum_i^n z_i{}^2\| \leq n\sum_i \mathbb{E}\Big[\|z_i{}^2\|\Big]$ holds for any random variable z_i, and 1.

The term B_2 will be bounded as follows:

$$
-\mu\Bigg\langle \nabla F^s(\mathbf{x}_S^t), \mathbb{E}\Bigg[\sum_{i\in[S_t]} \nabla F_i^s(\mathbf{x}_S^t)\Bigg]\Bigg\rangle \overset{(b_5)}{=} -n\mu\mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\|
\tag{A52}
$$

Note that equality (b_5) holds based on the definition 3.

The term B_3 will be bounded as below:

$$
\begin{aligned}
&\mathbb{E}\|\sum_{i\in[S_t]} \tilde{\nabla} F_i^s(\mathbf{x}_{S,i}^t)^2\| \overset{(b_6)}{\leq} \mathbb{E}\|\sum_{i\in[S_t]} \nabla F_i^s(\mathbf{x}_{S,i}^t)^2\| + \mathbb{E}\|\sum_{i\in[S_t]} (\tilde{\nabla} F_i^s(\mathbf{x}_{S,i}^t) - \nabla F_i^s(\mathbf{x}_{S,i}^t))^2\| \\
&\overset{(b_7)}{\leq} n\sigma_L^2 + \mathbb{E}\|\sum_{i=1}^{n} \nabla F_{q_i^t}^s(\mathbf{x}_{S,q_i^t}^t)^2\|
\end{aligned}
\tag{A53}
$$

Now, let's consider $\mathbf{t}_i = \nabla F_i^s(\mathbf{x}_{S,i}^t)$ following the ideas of [15] for this part:

$$
\begin{aligned}
&\mathbb{E}\left[\|\sum_{i=1}^{n} \nabla F_{q_i^t}^s(\mathbf{x}_{S,q_i^t}^t)^2\|\right] = \mathbb{E}\left[\|\sum_{i=1}^{n} \mathbf{t}_{q_i^t}{}^2\|\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} \|\mathbf{t}_{q_i^t}{}^2\| + \sum_{i \neq z, q_i^t, q_z^t \in [S_t]} \langle \mathbf{t}_{q_i^t}, \mathbf{t}_{q_z^t} \rangle\right] \\
&\overset{b_8}{=} \mathbb{E}\left[n\|\mathbf{t}_{q_1^t}{}^2\| + n(n-1)\langle \mathbf{t}_{q_1^t}, \mathbf{t}_{q_2^t} \rangle\right] \\
&= \frac{n}{m} \sum_{i \in [S]} \|\mathbf{t}_i{}^2\| + \frac{n(n-1)}{m^2} \sum_{i,z \in [S]} \langle \mathbf{t}_i, \mathbf{t}_z \rangle \\
&= \frac{n}{m} \sum_{i \in [S]} \|\mathbf{t}_i{}^2\| + \frac{n(n-1)}{m^2} \|\sum_{i \in [S]} \mathbf{t}_i{}^2\| \\
&\overset{b_9}{\leq} \frac{n^2}{m} \underbrace{\sum_{i \in [S]} \|\mathbf{t}_i{}^2\|}_{B_4} \qquad \text{(A54)}
\end{aligned}
$$

Note that the equality b_8 is due to independent sampling with replacement as outlined by [15]. The inequality b_9 follows from A8. We bound the term B_4 as follows:

$$
\begin{aligned}
&\sum_{i \in [S]} \|\mathbf{t}_i{}^2\| = \\
&= \sum_{i=0}^{m-1} \mathbb{E}\|\nabla F_i^s(\mathbf{x}_{S,i}^t)^2\| \\
&= \sum_{i=0}^{m-1} \mathbb{E}\|\nabla F_i^s(\mathbf{x}_{S,i}^t) - \nabla F^s(\mathbf{x}_{S,i}^t) + \nabla F^s(\mathbf{x}_{S,i}^t) - \nabla F^s(\mathbf{x}_S^t) + \nabla F^s(\mathbf{x}_S^t)^2\| \\
&\leq 3\sum_{i=0}^{m-1} \mathbb{E}\|\nabla F_i^s(\mathbf{x}_{S,i}^t) - \nabla F^s(\mathbf{x}_{S,i}^t)^2\| + 3\sum_{i=0}^{m-1} \mathbb{E}\|\nabla F^s(\mathbf{x}_{S,i}^t) - \nabla F^s(\mathbf{x}_S^t)^2\| + \\
&3\sum_{i=0}^{m-1} \mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| \\
&\overset{(b_{10})}{\leq} 3m\sigma_G^2 + 3L^2 \underbrace{\sum_{i=0}^{m-1} \mathbb{E}\|\mathbf{x}_{S,i}^t - \mathbf{x}_S^t{}^2\|}_{B_5} + 3m\mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| \qquad \text{(A55)}
\end{aligned}
$$

The inequality (b_{10}) holds due to Assumptions 3.

The term B_5 is bounded similar to [13, Lemma 3]:

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{x}_{S,i}^t - \mathbf{x}_S^t{}^2\| \overset{(b_{11})}{\leq} \mathbb{E}\|\mathbf{x}_{S,i-1}^t - \mathbf{x}_S^t - \mu\tilde{\nabla} F_{i-1}^s(\mathbf{x}_{S,i-1}^t)^2\| \\
&= \mathbb{E}\|\mathbf{x}_{S,i-1}^t - \mathbf{x}_S^t{}^2\| + \mathbb{E}\|\mu\tilde{\nabla} F_{i-1}^s(\mathbf{x}_{S,i-1}^t)^2\| + 2\left\langle \mathbf{x}_{S,i-1}^t - \mathbf{x}_S^t, -\mu\tilde{\nabla} F_{i-1}^s(\mathbf{x}_{S,i-1}^t) \right\rangle
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(b_{12})}{\leq} (1+\frac{1}{2m-1})\mathbb{E}\|\mathbf{x}^t_{S,i-1}-\mathbf{x}^{t}_S{}^2\| + (1+2m)\mathbb{E}\|\mu\tilde{\nabla} F^s_{i-1}(\mathbf{x}^t_{S,i-1})^2\| \\
&= (1+\frac{1}{2m-1})\mathbb{E}\|\mathbf{x}^t_{S,i-1}-\mathbf{x}^{t}_S{}^2\| + (1+2m)\mu^2\mathbb{E}\Bigg\|\tilde{\nabla} F^s_{i-1}(\mathbf{x}^t_{S,i-1}) - \nabla F^s_{i-1}(\mathbf{x}^t_{S,i-1}) + \\
&\nabla F^s_{i-1}(\mathbf{x}^t_{S,i-1}) - \nabla F^s_{i-1}(\mathbf{x}^t_S) + \nabla F^s_{i-1}(\mathbf{x}^t_S) - \nabla F^s(\mathbf{x}^t_S) + \nabla F^s(\mathbf{x}^t_S)\Bigg\|^2 \\
&\overset{(b_{13})}{\leq} (1+\frac{1}{2m-1})\mathbb{E}\|\mathbf{x}^t_{S,i-1}-\mathbf{x}^{t}_S{}^2\| + 4(1+2m)\mu^2\mathbb{E}\|\tilde{\nabla} F^s_{i-1}(\mathbf{x}^t_{S,i-1}) - \nabla F^s_{i-1}(\mathbf{x}^t_{S,i-1})^2\| + \\
&4(1+2m)\mu^2\mathbb{E}\|\nabla F^s_{i-1}(\mathbf{x}^t_{S,i-1}) - \nabla F^s_{i-1}(\mathbf{x}^t_S)^2\| + 4(1+2m)\mu^2\mathbb{E}\|\nabla F^s_{i-1}(\mathbf{x}^t_S\|) - \nabla F^s(\mathbf{x}^t_S)^2 \\
&+ 4(1+2m)\mu^2\mathbb{E}\|\nabla F^s(\mathbf{x}^t_S)^2\| \\
&\overset{(b_{14})}{\leq} (1+\frac{1}{2m-1} + 4(1+2m)\mu^2L^2)\mathbb{E}\|\mathbf{x}^t_{S,i-1}-\mathbf{x}^{t}_S{}^2\| + 4(1+2m)\mu^2(\sigma_L^2+\sigma_G^2) + \\
&4(1+2m)\mu^2\mathbb{E}\|\nabla F^s(\mathbf{x}^t_S)^2\|
\end{aligned}
$$

The A48 implies the inequality (b_{11}). The inequality (b_{12}) holds true based on $2\langle a,b\rangle \leq \frac{1}{n-1}\|a^2\| + n\|b^2\|$ for any two vectors a, b and positive number n. The inequality (b_{13}) follows from the previously mentioned fact at the inequality (b_3), and the inequality (b_{14}) is based on Assumptions 3 and 1.

Given $\mu \leq \frac{1}{2L(2m+1)}$ and by averaging over the clients, observe that:

$$
\begin{aligned}
&\frac{1}{m}\sum_{i=0}^{m-1}\mathbb{E}\|\mathbf{x}^t_{S,i}-\mathbf{x}^{t}_S{}^2\| \leq (1+\frac{4m}{4m^2-1})\frac{1}{m}\sum_{i=1}^{m-1}\mathbb{E}\|\mathbf{x}^t_{S,i-1}-\mathbf{x}^{t}_S{}^2\| + 4(1+2m)\mu^2(\sigma_L^2+\sigma_G^2) + \\
&4(1+2m)\mu^2\mathbb{E}\|\nabla F^s(\mathbf{x}^t_S)^2\| \\
&\leq (1+\frac{1}{m-1})\frac{1}{m}\sum_{i=1}^{m-1}\mathbb{E}\|\mathbf{x}^t_{S,i-1}-\mathbf{x}^{t}_S{}^2\| + 4(1+2m)\mu^2(\sigma_L^2+\sigma_G^2) + 4(1+2m)\mu^2\mathbb{E}\|\nabla F^s(\mathbf{x}^t_S)^2\|
\end{aligned}
\tag{A56}
$$

Unrolling the recursion, following [13, Lemma 3], it is inferred that:

$$
\begin{aligned}
&\frac{1}{m}\sum_{i=0}^{m-1}\mathbb{E}\|\mathbf{x}^t_{S,i}-\mathbf{x}^{t}_S{}^2\| \\
&\leq \sum_{j=0}^{m-1}(1+\frac{1}{m-1})^j\Big(4(1+2m)\mu^2(\sigma_L^2+\sigma_G^2) + 4(1+2m)\mu^2\mathbb{E}\|\nabla F^s(\mathbf{x}^t_S)^2\|\Big) \\
&\leq (m-1)\times\left[\left(1+\frac{1}{m-1}\right)^m - 1\right]\times\left[4(1+2m)\mu^2(\sigma_L^2+\sigma_G^2) + 4(1+2m)\mu^2\mathbb{E}\|\nabla F^s(\mathbf{x}^t_S)^2\|\right] \\
&\leq 16(m+2m^2)\mu^2(\sigma_L^2+\sigma_G^2) + 16(m+2m^2)\mu^2\mathbb{E}\|\nabla F^s(\mathbf{x}^t_S)^2\|
\end{aligned}
\tag{A57}
$$

Note that in the above inequality, $\left(1+\frac{1}{m-1}\right)^m - 1 \leq 4$ for $m > 1$.

Plugging A51, A52, A55 and A57 into A50, observe that

$$\begin{aligned}
\mathbb{E}\Big[F^s(\mathbf{x}_S^{t+l})\Big] &\le F^s(\mathbf{x}_S^t) + \frac{1}{2}\Big(16L^4(n^3+2n^4)\mu^3 + Ln\mu^2 + 48L^3n^2\mu^4(m+2m^2)\Big)\sigma_L^2 \\
&+ \frac{1}{2}\Big(16L^4(n^3+2n^4)\mu^3 + 3L\mu^2n^2 + 48L^3n^2\mu^4(m+2m^2)\Big)\sigma_G^2 \\
&+ \frac{1}{2}\Big(\mu - 2n\mu + 16L^4(n^3+2n^4)\mu^3 + 3L\mu^2n^2 + 48L^3n^2\mu^4(m+2m^2)\Big)\mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| \\
&\overset{(b_{15})}{\le} F^s(\mathbf{x}_S^t) + \frac{\mu(4-2m)}{2}\mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| + \frac{L\mu m^2}{2}\Big(14\sigma_L^2 + 14\sigma_G^2\Big)
\end{aligned} \tag{A58}$$

After simplifications, the inequality b_{15} holds as $\mu \le \frac{1}{8L^2m^2}$. Rearranging the terms, and summing over t, observe that:

$$\sum_t \mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| \le \frac{F^s(\mathbf{x}_S^0) - F^s(\mathbf{x}_S^*)}{\mu(m-2)} + \sum_t \frac{L\mu m^2}{m-2}\Big(7\sigma_L^2 + 7\sigma_G^2\Big) \tag{A59}$$

Assuming there are T global rounds overall,

$$\min_t \mathbb{E}\|\nabla F^s(\mathbf{x}_S^t)^2\| \le \frac{l\Big(F^s(\mathbf{x}_S^0) - F^s(\mathbf{x}_S^*)\Big)}{\mu(m-2)T} + \frac{L\mu m^2}{m-2}\Big(7\sigma_L^2 + 7\sigma_G^2\Big) \tag{A60}$$

This concludes the proof. □

References

1. McMahan, H.B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-Efficient Learning of Deep Networks from Decentralized Data, 2023, [arXiv:cs.LG/1602.05629].
2. Thapa, C.; Chamikara, M.A.P.; Camtepe, S. SplitFed: When Federated Learning Meets Split Learning. *CoRR* **2020**, *abs/2004.12088*, [2004.12088].
3. Gupta, O.; Raskar, R. Distributed learning of deep neural network over multiple agents. *Journal of Network and Computer Applications* **2018**, *116*, 1–8.
4. Mu, Y.; Shen, C. Communication and Storage Efficient Federated Split Learning. *arXiv preprint arXiv:2302.05599* **2023**.
5. Kim, M.; DeRieux, A.; Saad, W. A bargaining game for personalized, energy efficient split learning over wireless networks. 2023 IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2023, pp. 1–6.
6. Li, Y.; Lyu, X. Convergence Analysis of Sequential Split Learning on Heterogeneous Data. *arXiv preprint arXiv:2302.01633* **2023**.
7. Liao, Y.; Xu, Y.; Xu, H.; Yao, Z.; Wang, L.; Qiao, C. Accelerating federated learning with data and model parallelism in edge computing. *IEEE/ACM Transactions on Networking* **2023**.
8. Han, D.J.; Bhatti, H.I.; Lee, J.; Moon, J. Accelerating federated learning with split learning on locally generated losses. ICML 2021 workshop on federated learning for user privacy and data confidentiality. ICML Board, 2021.
9. Belilovsky, E.; Eickenberg, M.; Oyallon, E. Decoupled greedy learning of cnns. International Conference on Machine Learning. PMLR, 2020, pp. 736–745.
10. Jaderberg, M.; Czarnecki, W.M.; Osindero, S.; Vinyals, O.; Graves, A.; Silver, D.; Kavukcuoglu, K. Decoupled neural interfaces using synthetic gradients. International conference on machine learning. PMLR, 2017, pp. 1627–1635.
11. Ghadimi, S.; Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization* **2013**, *23*, 2341–2368.

12. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; others. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning* **2021**, *14*, 1–210.
13. Reddi, S.; Charles, Z.; Zaheer, M.; Garrett, Z.; Rush, K.; Konečnỳ, J.; Kumar, S.; McMahan, H.B. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295* **2020**.
14. Reisizadeh, A.; Mokhtari, A.; Hassani, H.; Jadbabaie, A.; Pedarsani, R. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. International conference on artificial intelligence and statistics. PMLR, 2020, pp. 2021–2031.
15. Yang, H.; Fang, M.; Liu, J. Achieving linear speedup with partial worker participation in non-iid federated learning. *arXiv preprint arXiv:2101.11203* **2021**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.