Local Averaging Helps: Hierarchical Federated Learning and Convergence Analysis

Jiayi Wang*, Shiqiang Wang†, Rong-Rong Chen*, Mingyue Ji*

Abstract

Federated learning is an effective approach to realize collaborative learning among edge devices without exchanging raw data. In practice, these devices may connect to local hubs which are then connected to the global server (aggregator). Due to the (possibly limited) computation capability of these local hubs, it is reasonable to assume that they can perform simple averaging operations. A natural question is whether such local averaging is beneficial under different system parameters and how much gain can be obtained compared to the case without such averaging. In this paper, we study hierarchical federated learning with stochastic gradient descent (HF-SGD) and conduct a thorough theoretical analysis to analyze its convergence behavior. The analysis demonstrates the impact of local averaging precisely as a function of system parameters. Due to the higher communication cost of global averaging, a strategy of decreasing the global averaging frequency and increasing the local averaging frequency is proposed. Experiments validate the proposed theoretical analysis and the advantages of hierarchical federated learning.

1 Introduction

Federated learning (FL) is an emerging distributed machine learning technique that enables learning at a large scale from decentralized data generated by client devices [8, 11, 14, 19, 21, 23]. In particular, it has been proposed to be applied to wireless edge and mobile networks, where a massive number of devices collaborate together to perform a learning task without exchanging raw data [7, 13, 14, 20]. Compared to conventional machine learning techniques where data need to be transmitted to and stored at a central location, the benefits of FL include improved data privacy and reduced communication and storage requirements.

In a typical FL system, a server in the cloud acts as a parameter aggregator that receives iterative updates from clients distributed over a large geographical area. After receiving the updates, the server aggregates (averages) the parameters and sends the result back to clients for the next iteration. The challenge with this approach is that it includes a lot of communication over wide-area network (WAN), which is prone to unpredictable latency and packet loss especially in areas with poor network connections. This can cause poorly connected areas to be unfairly treated in the model training process, so that the FL result is in favor of data originated from well-connected clients.

To resolve this problem, we present a hierarchical FL technique in this paper, where two “levels” of parameter aggregation functionalities are distributed across the networked system, as shown in Figure 1. In particular, local averaging can be performed at a networked component that is in close proximity to...
a certain group of clients. As local communication bandwidth is usually much higher than WAN bandwidth, this approach can alleviate network congestion, increase the communication robustness of clients, and also reduce the overall ingress traffic at the server. From a system perspective, recent developments of edge computing and network function virtualization (NFV) provide computational capability to various components (e.g., edge nodes, gateways, routers) in future communication networks [1]. Local parameter averaging can be performed on such networked components.

To support hierarchical FL with local averaging, a fundamental question is: how does local averaging affect FL convergence and what is the impact of data distribution at FL clients on this process? In this paper, we address this question by conducting a thorough theoretical analysis with empirical validations. Our main contributions are summarized as follows.

- We present a hierarchical FL method based on stochastic gradient descent (SGD) (referred to as HF-SGD) that leverages local averaging within groups of clients.
- We theoretically analyze the convergence bound of HF-SGD for both fixed client grouping and uniformly random client grouping, with non-independent/non-identically distributed (non-IID) data across the clients and groups. This is the first work to analyze the effect of client grouping by considering both local and global divergences in HF-SGD. Our analytical bound clearly demonstrates the advantage of HF-SGD over single-level federated SGD (SF-SGD).
- Our theoretical results lead to useful insights such as how local and global averaging periods (intervals) and data heterogeneity affect the convergence of HF-SGD. These are validated with experiments on real datasets. Our experiments show that HF-SGD brings the benefit of accelerated learning convergence while reducing the communication cost of global averaging.

2 Related Work

FL was first proposed by McMahan et al. [19] as a federated averaging (FedAvg) algorithm. It was later shown that FedAvg converges with non-IID client data [15,24]. The non-IID data distribution at clients is one of the major challenges in FL, which is also known as data heterogeneity.

To improve the communication efficiency of FL, methods for adapting the parameter averaging period (interval) for efficient learning were proposed [22,23]. Model/gradient sparsification and compression techniques were also applied to FL, which reduces the amount of information exchanged between clients and the server while maintaining a similar model accuracy [6,9,10,11]. In addition, partial client participation can further reduce the computation and communication overhead [2], which, however, still
requires the participating clients to communicate with the server directly and the convergence rate is inversely proportional to the number of participating clients \([15]\). These techniques are orthogonal to our work and may be applied together with our proposed hierarchical FL.

By moving parameter aggregation from the server down to the clients, decentralized SGD was proposed by Lian et al. [16, 17]. However, it does not address the common FL scenario where an aggregation is only performed once after multiple local iterations for communication efficiency, and does not support a hierarchical setup with aggregation at local hubs either. Some recent works have considered similar settings as hierarchical FL. Chen et al. [5] studied sequential updates from each group, which is different from our HF-SGD method that allows parallel updates with two levels of (local and global) aggregation. Independent and identically distributed (IID) data was considered in the work by Castiglia et al. [4], which does not capture the impact of data heterogeneity. The result by Liu et al. [18] has a bound that is exponential in the parameter aggregation period (interval) and can become loose if the period is large.

In this paper, different from the above existing works, we derive the convergence bound of hierarchical FL (more specifically, HF-SGD) that captures the impact of data heterogeneity within and across groups of clients, based on which useful insights are observed. In addition, our bound is tight and degenerates to the best known results for SF-SGD and non-federated (i.e., centralized) SGD. Our analysis is challenging due to the complex interplay between local aggregation (within a group) and global aggregation (across groups), data heterogeneity, and convergence.

In the remainder of this paper, Section 3 describes the overall procedure of HF-SGD and Section 4 gives its convergence analysis. Strategies for reducing communication cost are discussed in Section 5. The experimental results are presented in Section 6. Finally, Section 7 draws conclusion.

### 3 Hierarchical Federated Stochastic Gradient Descent (HF-SGD)

In classical FedAvg [19], a set of clients \( V := \{1, 2, \ldots, n\} \) collaboratively minimize the overall empirical risk in an iterative manner. After a number of local SGD iterations (referred to as a local period) computed on each client’s local data, the resulting model parameters are sent to the server, which performs an averaging operation to obtain an intermediate global model with parameter vector \( \bar{w} \). Then, \( \bar{w} \) is sent to all clients, and the FL process continues until convergence. As mentioned earlier, we refer to this classical FedAvg method as SF-SGD, to differentiate from our HF-SGD technique.

In our hierarchical FL system (see Figure 1), all clients are partitioned into \( N \) groups \( V_1, V_2, \ldots, V_N \). The number of clients in each group is denoted by \( n_i := |V_i| (i = 1, 2, \ldots, N) \). The grouping can be a natural result of clients connecting to different local hubs, where those clients that are connected to a common hub can be considered as one group. In this case, grouping is fixed based on the network topology. In another example, grouping can be based on the dynamic configuration of parameter aggregation functions and routing in an NFV-supported system, which is largely configurable at a logical level and allows strategies such as random grouping. In this paper, our main goal is to characterize the impact of grouping and show its advantages. For a given grouping, we can write the objective function as:

\[
\min_{w \in \mathbb{R}^d} f(w) := \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j \in V_i} F_j(w) \tag{1}
\]

where \( w \in \mathbb{R}^d \) is the model parameter vector and \( F_j(w) := \sum_{s \in D_j} \ell(w, s) \) is the empirical risk function defined over the local dataset \( D_j \) at client \( j \) (\( \ell(w, s) \) is the loss of data sample \( s \)). In this paper, we consider a general \( F_j(\cdot) \) that is non-convex. Then, for each group \( i \), we define the intra-group objective function as:

\[
f_i(w) := \frac{1}{n_i} \sum_{j \in V_i} F_j(w). \tag{2}
\]

\( ^2 \)In general, the results of this paper can be extended to the case of weighted averaging as well.
During each local iteration \( t \), each client \( j \) updates its own model using SGD:

\[
\mathbf{w}_{j}^{t+1} = \mathbf{w}_{j}^{t} - \gamma \mathbf{g}(\mathbf{w}_{j}^{t}, \zeta_{j}^{t})
\]

where \( \gamma \) is the learning rate, \( \mathbf{g}(\mathbf{w}_{j}^{t}, \zeta_{j}^{t}) \) is the stochastic gradient of \( F_{j}(\mathbf{w}) \), and \( \zeta_{j}^{t} \) represents random data samples from the local dataset \( \mathcal{D}_{j} \) at client \( j \). We assume that \( \mathbb{E}_{\zeta_{j}^{t} \sim \mathcal{D}_{j}}[\mathbf{g}(\mathbf{w}_{j}^{t}, \zeta_{j}^{t})] = \nabla F_{j}(\mathbf{w}_{j}^{t}) \).

Instead of aggregating all local models at the global aggregator (server) at the end of a local period as in SF-SGD, models are first averaged within group \( i \) (\( i \in \{1, 2, ..., N\} \)) after every \( I_{i} \) local iterations. In particular, at local iteration \( t \in \{I_{i}, 2I_{i}, 3I_{i}, \ldots\} \), we compute \( \bar{\mathbf{w}}_{i}^{t} := \frac{1}{n_{i}} \sum_{j \in \mathcal{V}_{i}} \mathbf{w}_{j}^{t} \). This can be done at a local aggregator (e.g., a computational component in close proximity) for group \( i \). After several rounds of “intra-group” averagings, models from the \( N \) groups are averaged globally. Let a global aggregation be performed for every \( G \) local iterations. Then, at local iteration \( t \), we have \( \bar{\mathbf{w}}^{t} := \frac{1}{N} \sum_{i=1}^{N} \bar{\mathbf{w}}_{i}^{t} \) for \( t = G, 2G, 3G, \ldots \). Note that we assume that all the clients perform synchronous updates and we let \( G \) be a common multiple of \( \{I_{1}, \ldots, I_{N}\} \). Therefore, FL is conducted both at the local level within a group and the global level across groups. This design is referred to as HF-SGD, which is summarized in Algorithm 1

### 4 Convergence Analysis

We begin with a description of the assumptions made in our convergence analysis. Then, we present our results for HF-SGD with fixed client groupings. These will be followed by our results for HF-SGD with random grouping. The analysis presented in this work characterizes the convergence behavior of HF-SGD and clearly demonstrates its advantage over SF-SGD.

#### 4.1 Assumptions

**Assumption 1.** In this paper, we make the following assumptions.

- **a) Lipschitz gradient**

\[
\|\nabla F_{j}(\mathbf{w}) - \nabla F_{j}(\mathbf{w}')\| \leq L\|\mathbf{w} - \mathbf{w}'\|, \forall j, \mathbf{w}, \mathbf{w}'
\]

where \( L \) is some positive constant.
b) **Bounded variance**

$$\mathbb{E}_{z_j \sim D_j} \| g(w; z_j) - \nabla F_j(w) \|^2 \leq \sigma^2, \forall j, w$$  \hspace{1cm} (5)

where $D_j$ is the dataset at client $j \in \mathcal{V}$.

c) **Bounded global divergence (HF-SGD)**

$$\frac{1}{N} \sum_{i=1}^{N} \| \nabla f_i(w) - \nabla f(w) \|^2 \leq \epsilon^2, \forall i, w$$  \hspace{1cm} (6)

d) **Bounded local divergence (HF-SGD)**

$$\frac{1}{n_i} \sum_{j \in \mathcal{V}_i} \| \nabla F_j(w) - \nabla f_i(w) \|^2 \leq \epsilon_i^2, \forall i, w$$  \hspace{1cm} (7)

Note that the Lipschitz gradient assumption also applies to the group objective $f_i(w)$ and overall objective $f(w)$. For example, $\| \nabla f_i(w) - \nabla f_i(w') \| = \| \frac{1}{n} \sum_{j \in \mathcal{V}_i} \nabla F_j(w) - \frac{1}{n} \sum_{j \in \mathcal{V}_i} \nabla F_j(w') \| \leq \frac{1}{n_i} \sum_{j \in \mathcal{V}_i} \| \nabla F_j(w) - \nabla F_j(w') \| \leq L \| w - w' \|$.

For the special case of $N = 1$, all $n$ clients form a single group and HF-SGD reduces to SF-SGD. In this case, we can re-write (7) as the following.

**Assumption 2. Bounded divergence (SF-SGD)**

$$\frac{1}{n} \sum_{j \in \mathcal{V}} \| \nabla F_j(w) - \nabla f(w) \|^2 \leq \tilde{\epsilon}^2, \forall w.$$  \hspace{1cm} (8)

To characterize the impact of client grouping in HF-SGD, instead of assuming (8), we assume (6) and (7) for bounded global and local divergences, respectively. In particular, the global divergence corresponds to the relationship between the gradients of intra-group objectives (defined on clients’ data within each group) and the gradient of the overall objective (defined on all clients’ data). A larger global divergence means a more heterogeneous data distribution between groups. Similarly, the local divergence $\epsilon_i$ characterizes the data heterogeneity of clients within each group $i$.

### 4.2 HF-SGD with Fixed Client Grouping

In the following, we will present our main convergence results when the client grouping is fixed.

**Theorem 1.** Consider the problem in (1). For any fixed client grouping that satisfies Assumption 2 if the learning rate in Algorithm 1 satisfies $\gamma < \frac{1}{2 \sqrt{3GL}}$, then for any $T \geq 1$, we have\footnote{Since $F_j(\cdot), f_i(\cdot), f(\cdot)$ are non-convex, we focus on characterizing the upper bound of $\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \| \nabla f(\bar{w}^t) \|$.}

\[
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \| \nabla f(\bar{w}^t) \|^2 \leq 2 (f^0 - f^*) + \frac{\gamma L \sigma^2}{\gamma T} + \frac{N^2 \sum_{i=1}^{N} \frac{1}{n_i}}{N} + \frac{4\gamma^2 L^2 \left( \left( 1 - \frac{1}{n} \right) \left( \frac{1}{n} \sum_{i=1}^{N} \frac{1}{n_i} \right) 2G\sigma^2 + 3G^2\epsilon^2 \right)}{1 - 12\gamma^2 G^2 L^2} \quad \text{(9a)}
\]

\[
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \| \nabla f(\bar{w}^t) \|^2 \leq 2 (f^0 - f^*) + \frac{\gamma L \sigma^2}{\gamma T} + \frac{N^2 \sum_{i=1}^{N} \frac{1}{n_i}}{N} + \frac{4\gamma^2 L^2 \left( \left( 1 - \frac{1}{n} \right) \left( \frac{1}{n} \sum_{i=1}^{N} \frac{1}{n_i} \right) 2G\sigma^2 + 3G^2\epsilon^2 \right)}{1 - 12\gamma^2 G^2 L^2} \quad \text{(9b)}
\]
\[
\frac{8\gamma^2G^2L^2}{1 - 12\gamma^2G^2L^2} \cdot \frac{4\gamma^2L^2}{1 - 12\gamma^2I_{\text{max}}^2L^2} \cdot \frac{1}{N} \sum_{i=1}^{N} \left[(1 - \frac{1}{n_i})I_i \sigma^2 + 3I_i^2\epsilon_i^2\right]
\]

where \(I_{\text{max}} := \max_i I_i\).

All missing proofs in this paper can be found in the appendix.

Note that the bound \((9a)–(9c)\) can be partitioned into three parts. The two terms in \((9a)\) correspond to the impact of non-federated SGD, which means that if we let \(N = n = 1\), then \(\epsilon_i = \epsilon = 0\) according to their definitions and we have \(\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(\bar{w}^t)\|^2 \leq \frac{2(f^0 - f^*)}{\gamma T} + \gamma L\sigma^2\). This is the same as the non-federated SGD result by Bottou et al. \[3\]. It shows that our bound reduces to the non-federated SGD case when \(N = n = 1\). The term in \((9b)\) shows the impact of global (cross-group) averaging. The term in \((9c)\) mainly corresponds to the impact of local (intra-group) averaging. Note that \(G\) also appears in \((9c)\). It can be interpreted as follows. When \(I_i\)s are fixed and \(G\) increases, the bound in \((9c)\) becomes larger since a greater \(G\) reduces the global averaging frequency.

In \((9b)\) and \((9c)\), it can be seen that the coefficients in front of the SGD noise \(\sigma^2\) involve \(G\) and \(I_i\), while the coefficients in front of the divergences \(\epsilon\) and \(\epsilon_i\) involve \(G^2\) and \(I_i^2\). This means that the divergences may play a more important role compared to \(\sigma^2\) on the convergence rate. One implication from this observation is that if \(G^2\) is fixed, then it is desirable to choose a smaller \(I_i\) for a bigger \(\epsilon_i\), and a larger \(I_i\) for a smaller \(\epsilon_i\) (in a way that the bound \((9a)–(9c)\) remains unchanged), which follows the intuition.

When \(\frac{8\gamma^2G^2L^2}{1 - 12\gamma^2G^2L^2} < 1\), we can obtain the following corollary directly from Theorem 1, which gives a slightly looser but simpler upper bound of \(\sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(\bar{w}^t)\|^2\).

**Corollary 1.** Consider the problem in \((7)\). For any fixed client grouping that satisfies Assumption \[7\] if the learning rate in Algorithm 1 satisfies \(\gamma \leq \frac{1}{2\sqrt{6GL}}\), then for any \(T \geq 1\), we have

\[
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(\bar{w}^t)\|^2 \leq \frac{2(f^0 - f^*)}{\gamma T} + \gamma L\sigma^2 \sum_{i=1}^{N} \frac{1}{n_i} + 2C\gamma^2GL^2 \left(1 - \frac{1}{N}\right) \left(\frac{\sigma^2}{N} \sum_{i=1}^{N} \frac{1}{n_i}\right) + 3C\gamma^2G^2L^2\epsilon^2
\]

\[
+ \frac{2C\gamma^2G^2L^2\sigma^2}{N} \sum_{i=1}^{N} \left(1 - \frac{1}{n_i}\right) I_i + \frac{3C\gamma^2G^2L^2}{N} \sum_{i=1}^{N} I_i^2\epsilon_i^2,
\]

where \(C = 40/3\).

**Remark.** From Corollary \[1\], let \(\gamma = 1/\sqrt{T \left(\frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i}\right)}\) with \(T \geq \sqrt{\frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i}} / 24G^2L^2\).

when \(T\) is sufficiently large, it can be seen that

\[
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(\bar{w}^t)\|^2 = O\left(\sqrt{\frac{1}{T} \sum_{i=1}^{N} \frac{1}{n_i}}\right) + O\left(\frac{1}{T}\right).
\]

If \(n_i = n/N\) (equal number of clients per group), then we can obtain

\[
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(\bar{w}^t)\|^2 = O\left(\frac{1}{\sqrt{nT}}\right) + O\left(\frac{1}{T}\right),
\]

which achieves a linear speedup compared to the case of non-federated SGD, which aligns with the observation by Yu et al. \[24\].
In the following corollary, we will show that (10) of Corollary 1 can be reduced to the case of SF-SGD. We use $P$ to denote the aggregation period of SF-SGD.

**Corollary 2. (Degenerate to SF-SGD case)** Let $N = 1$ and $\gamma \leq \frac{1}{2\sqrt{6}PL}$, from (10), we obtain

$$\frac{1}{T} \sum_{t=0}^{T-1} E\|\nabla f(\bar{w}_t)\|^2 \leq \frac{2(f^0 - f^*)}{\gamma T} + \frac{\gamma L\sigma^2}{n} + 2C\gamma^2 L^2 \sigma^2 \left( 1 - \frac{1}{n} \right) P + 3C\gamma^2 L^2 P^2 \tilde{\epsilon}^2$$

$$= O\left( \frac{1}{\gamma T} \right) + O\left( \frac{\gamma^2 \sigma^2}{n} \right) + O\left( \gamma^2 P^2 \left( 1 - \frac{1}{n} \right) \right) + O\left( \gamma^2 P^2 \tilde{\epsilon}^2 \right). \quad (13)$$

The proof of Corollary 2 is straightforward by letting $N = 1$, $n_1 = n$, and $I = P$ in (10) and using (8).

While (13) is similar to the bound by Yu et al. [24], we note that the third term $O\left( \gamma^2 P^2 \left( 1 - \frac{1}{n} \right) \right)$ in the last equality in (13) has an additional term of $\left( 1 - \frac{1}{n} \right)$ compared to the bound in Yu et al. [24]. This term can potentially make our bound tighter. Another important observation is that the techniques used to obtain this term is the key to make our HF-SGD bound (10) degenerate to the SF-SGD and non-federated SGD cases.

### 4.3 HF-SGD with Random Grouping and Comparison to SF-SGD

Built upon the results of Theorem 1 for a fixed client grouping, we further analyze the convergence performance of HF-SGD under the assumption of random grouping. The main technique used here is to first characterize the local and global divergence of HF-SGD under random grouping (see Lemmas 1 and 2 below) and relate them to the divergence of SF-SGD defined in (8). An upper bound on the averaged performance of HF-SGD is presented in Theorem 2 below, which demonstrates the advantage of HF-SGD over SF-SGD under random grouping.

For client grouping, we consider all possible grouping strategies with the constraint that $n_i = n/N \ (\forall i)$. Then, we uniformly select one grouping strategy at random. Let the random variable $S$ denote the uniformly random grouping strategy. This means that each realization of $S$ corresponds to one grouping realization. Based on the uniformly random grouping strategy $S$, in the following, we will show that under (8) in Assumption 1 we can explicitly compute the average global and local divergences, which can be considered as a counterpart of (6) and (7) in Assumption 1 as a function of $\tilde{\epsilon}$, $n$, and $N$.

Consider a random client $k$ in group $i$, we can obtain that $E_S[\nabla F_k(x)] = \nabla f(x)$, where the expectation $E_S$ is over all possible grouping strategies. These lead to the following two lemmas.

**Lemma 1.** Using the uniformly random grouping strategy $S$, the average global divergence is given by

$$E_S \left[ \frac{1}{N} \sum_{i=1}^{N} \left\| \nabla f(w) - \frac{N}{n} \sum_{k \in V_i} \nabla F_k(w) \right\|^2 \right] \leq \left( \frac{N-1}{n-1} \right) \hat{\epsilon}^2,$$ \quad (14)

where $\hat{\epsilon}$ is given in (9).

**Lemma 2.** Using the uniformly random grouping strategy $S$, the average local divergence is given by

$$E_S \left[ \frac{N}{n} \sum_{k \in V_i} \left\| \nabla f_k(w) - \frac{N}{n} \sum_{k' \in V_i} \nabla F_k'(w) \right\|^2 \right] \leq \left( 1 - \frac{N-1}{n-1} \right) \hat{\epsilon}^2,$$ \quad (15)

where $\hat{\epsilon}$ is given in (9).

Lemmas 1 and 2 show an interesting behavior of grouping. If we add the upper bounds of the average global divergence (14) and the average local divergence (15), then it equals to $\epsilon$, which is independent of
other system parameters. In Theorem 2, we will show that grouping can effectively “break” \( \tilde{\epsilon} \) into two parts, which are “modulated” by functions of \( G \) and \( I_i \), respectively.

For simplicity, we let each group have the same local period \( I_i = I, \forall i \). Then, we can obtain the following.

**Theorem 2.** Using the uniformly random grouping strategy \( S \), let \( \gamma \leq \frac{1}{2 \sqrt{6GL}} \), then we have

\[
\mathbb{E}_S \left[ \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \| \nabla f(\bar{w}_t) \|^2 \right] \leq \frac{2(f^0 - f^*)}{\gamma T} + \frac{\gamma L \sigma^2}{n} + 2C \gamma^2 L^2 \left[ \left( \frac{N-1}{n} \right) G + \left( 1 - \frac{N}{n} \right) I \right] \sigma^2
+ 3C \gamma^2 L^2 \left[ \left( \frac{N-1}{n-1} \right) G^2 + \left( 1 - \frac{N-1}{n-1} \right) I^2 \right] \tilde{\epsilon}^2,
\]

where \( C = 40/3 \).

The proof follows by substituting results of the expected divergence in Lemmas 1 and 2 to the convergence bound of Corollary 1. Without additional assumptions, the bound (16) can also degenerate to SF-SGD by setting \( N = 1 \) and letting \( I = P \).

**Remark.** Theorem 2 can be interpreted as follows. First, (16) explicitly shows the impact of grouping on the convergence rate of HF-SGD. For example, in the last term in (16), as mentioned earlier, the grouping approach “breaks” the divergence \( \tilde{\epsilon} \) into two parts, whose weights are determined by the choice of \( N \), \( G \), and \( I \). One implication is that since \( G^2 > I^2 \), it is preferable to have a smaller number of groups for a large \( G \). In addition, we note that the performance of HF-SGD can be sandwiched by the performance of two SF-SGDs. In order to see this, we consider the following three scenarios: 1) SF-SGD with averaging period \( P = G \); 2) SF-SGD with averaging period \( P = I \), and 3) HF-SGD with local averaging period \( I \) and global averaging period \( G \). We let them all start with the same \( \bar{w}_0 \) and choose learning rate \( \gamma \) such that \( \gamma \leq \frac{1}{2 \sqrt{6GL}} \). We can see that all three cases have the same first two terms in (16). However, for the third and fourth terms in (16), we have

\[
\left( 1 - \frac{1}{n} \right) I \leq \left( \frac{N-1}{n} \right) G + \left( 1 - \frac{N}{n} \right) I \leq \left( 1 - \frac{1}{n} \right) G,
\]

\[
I^2 \leq \left( \frac{N-1}{n-1} \right) G^2 + \left( 1 - \frac{N-1}{n-1} \right) I^2 \leq G^2.
\]

Eqs. (17) and (18) show that performance of HF-SGD can be lower and upper bounded by SF-SGD with averaging periods of \( P = I \) and \( P = G \), respectively. The grouping approach can explicitly characterize how much the convergence bound moves towards the best case, i.e., SF-SGD with \( P = I \). Note, however, that SF-SGD with \( P = I \) incurs the highest global communication cost, so grouping can adjust the trade-off between convergence and communication cost.

Although the discussion above is based on Theorem 2, which is for uniformly random grouping, our experiments in Section 6 show that this property of lower and upper bounded by SF-SGD also applies to fixed grouping in both IID and Non-IID cases.

## 5 Extension to Reducing Communication Cost

We introduce two approaches to reduce communication cost based on our theoretical results above. Later, we will provide experimental results in Section 6.

**Using more local aggregations.** In FL settings, global averaging is often more expensive than local averaging, as explained in Section 1. In this case, we can reduce the communication cost by decreasing
the frequency of global averaging (i.e., increasing $G$) while increasing the frequency of local averaging (i.e., decreasing $I$). Our theoretical analysis shows that this approach does not necessarily degrade the performance of HF-SGD. In some cases, e.g., if $I$ is sufficiently small, then the convergence performance of HF-SGD can actually be improved. To see this, consider the last two terms of (16). Suppose $G = mI$, $m = 1, 2, 3, \ldots$. For a non-trivial client grouping, i.e., $1 < N < n$, if we increase $G$ to $G' = lG$, $1 < l < \frac{\sqrt{n\gamma}}{N} + 1$ and decrease $I$ to $I' = qI$, $q \leq \frac{4 - m^2(l^2 - 1)}{n - N}$, then one can show that the bound (16) using $G'$ and $I'$ can be lower than or equal to that using $G$ and $I$. A similar behavior can also be seen for fixed grouping, as shown in the experimental results in Section 6.

Adjusting grouping strategy. Recall the divergence-related terms in Corollary 1, i.e., $3C\gamma^2L^2(G^2\epsilon^2 + \frac{1}{N} \sum_{i=1}^{N} I_i^2 \epsilon_i^2)$. Since $G^2$ is typically greater than $\frac{1}{N} \sum_{i=1}^{N} I_i^2$, decreasing the global divergence $\epsilon$ is more effective than decreasing the local divergence $\epsilon_i$ in order to tighten the convergence upper bound. If the knowledge of data distribution of all clients is available, we can adjust the grouping strategy to achieve a smaller global divergence. For example, we can make the data from each group distributed like a subset sampled uniformly from the union of all clients’ datasets, which is referred to as “group-IID”. In this way, we can increase $G$, and hence, reduce the global communication cost without affecting the convergence performance of HF-SGD.

6 Experiments

In this section, we validate our theoretical results with experiments to compare the performance of HF-SGD with SF-SGD. We also conduct experiments using the two approaches in Section 5 to reduce communication cost. The implementation is based on Pytorch.

6.1 Setup

Dataset and models. In our experiments, we train VGG-11 for the image classification tasks over the CIFAR-10 dataset which has 10 image classes [12], where we refer to the class labels as $\{0, 1, \ldots, 9\}$ for simplicity. In all experiments, we set the learning rate as $\gamma = 0.02$ and the SGD mini-batch size as 20.

Data separation. We implement four cases of data distribution. In the IID case, data are uniformly partitioned into 10 clients. In the non-IID case, there are 10 clients and data on each client are from only one image class. For example, the $j$-th client only has data from the $(j - 1)$-th class ($j = 1, 2, \ldots, 10$). Then, we also consider the data distribution across groups, where each group has 5 clients and the data among clients within each group is non-IID distributed. For the “group-IID” case, the 5 clients in each group have labels $\{0, 1\}, \{2, 3\}, \{4, 5\}, \{6, 7\}, \{8, 9\}$, respectively. In the “group-non-IID” case, clients in the first group have labels $\{0, 1\}, \{2, 3\}, \{4, 5\}, \{6, 7\}, \{8, 9\}$, respectively, while clients in the second group have labels $\{4, 5\}, \{6, 7\}, \{8, 9\}, \{6, 7\}, \{8, 9\}$, respectively.

6.2 Results

In all plots, curves with $P$ are for SF-SGD and curves with $G$, $I$, and $N$ are for HF-SGD.

Figure 2a shows the results of the IID case. In this case, we separate clients into two groups, each group with 5 clients. It can be seen that grouping has limited impact on the convergence when data on clients are IID. This is consistent with our theoretical results in Theorem 2, since IID implies that $\epsilon^2$ is very small. So the gain from decreasing $I$ is small. Nonetheless, the performances of different schemes align with Theorem 2. To illustrate this, we list the number of local iterations when first achieving an 80% accuracy in Table 1. We observe that the performances of HF-SGD with global period $G$ and local period $I$ are always between SF-SGDs with period $P = G$ and with period $P = I$. Note in this case, HF-SGD with $G = 100, I = 10$ performs only slightly better than SF-SGD with $P = 20$. This is because we only have two groups and 10 clients in total. In Theorem 2, the coefficient before $G^2$ is only $\frac{1}{2}$ in this case, which may reduce the impact of $G$.
Figure 2 shows the results of the non-IID scenario. For HF-SGD with \( N = 2 \), we partition 10 clients into two groups. In Group 1, there are only data with labels \{0, 1, 2, 3, 4\} while Group 2 only contains data with labels \{5, 6, 7, 8, 9\}. In this case, the benefit of grouping clients is significant. Comparing HF-SGD with \( G = 50, I = 10 \), and SF-SGD with \( P = 50 \), we see that by performing 4 local aggregations before global aggregation, HF-SGD can have a much improved convergence performance over SF-SGD. This is because \( \tilde{\epsilon}^2 \) is relatively large in the non-IID case. According to Theorem 2, by decreasing the value of \( \left( \frac{N-1}{n-1} \right) G^2 + \left( 1 - \frac{N-1}{n} \right) I^2 \) slightly, the convergence bound can be significantly reduced. In addition, we can also observe that the performance of HF-SGD with global period \( G \) and local period \( I \) is still between that of SF-SGD with \( P = G \) and with \( P = I \). Moreover, in Figure 2b, we plot a 5-group case where we partition the same set of clients into \( N = 5 \) groups. Each group has \( K = 2 \) clients. We see that the performance in this case is worse than the case with only two groups. This is because a larger \( N \) will make the term \( \frac{N-1}{n-1} \) larger so the coefficient for \( G \) becomes larger.

In Figure 3, we show the effects of grouping with group-IID strategy. In the group-IID case, data in each group are from all 10 labels. In group-non-IID case, data in each group are only from 6 labels (see Section 6.1). From (10) in Corollary 1, it can be seen that the bound depends on \( G^2 \tilde{\epsilon}^2 \). Hence, reducing \( \tilde{\epsilon}^2 \) can significantly improve the convergence performance since \( G^2 \) is desired to be large in general. Figure 3 also shows that the group-IID case performs as well as the group-non-IID case after reducing \( I \) by half.

In Figure 4, we show that decreasing \( I \) while increasing \( G \) can keep the performance while reducing communication cost. We see that HF-SGD with \( G = 200, I = 5 \) performs similarly to SF-SGD with \( P = 10 \). That is, by allowing more local aggregations, HF-SGD can reduce the number of global communication by 95\% while maintaining a similar performance to SF-SGD.

### 7 Conclusion

This paper has studied HF-SGD with two-level parameter aggregations. Based on this framework, we have successfully answered important questions on how local averaging affects FL convergence and what is the impact of data distribution at clients on the learning process. In particular, we have provided
a thorough theoretical analysis on the convergence of HF-SGD over non-IID data, under both fixed and random client grouping. Explicit comparisons of the convergence rate have been established for HF-SGD and SF-SGD, based on a novel analysis of the local and global divergences. Our theoretical analysis provides valuable insights into the design of practical HF-SGD systems, including the choice of global and local averaging periods. Different grouping strategies are considered to best utilize the divergence measures to reduce communication costs while accelerating learning convergence. Future work could analyze the effect of partial client participation in hierarchical FL and extend to more than two levels of parameter aggregation.

References

[1] Sophia Antipolis. ETSI multi-access edge computing builds on NFV and network slicing, 2020. URL https://www.etsi.org/newsroom/press-releases/1700-2020-01-etsi-multi-access-edge-computing-builds-on-nfv-and-network-slicing

[2] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečný, Stefano Mazzocchi, H Brendan McMahan, et al. Towards federated learning at scale: System design. In SysML 2019, 2019. URL https://arxiv.org/abs/1902.01046
[3] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[4] Timothy Castiglia, Anirban Das, and Stacy Patterson. Multi-level local sgd for heterogeneous hierarchical networks. *arXiv preprint arXiv:2007.13819*, 2020.

[5] Cheng Chen, Ziyi Chen, Yi Zhou, and Bhavya Kailkhura. Fedcluster: Boosting the convergence of federated learning via cluster-cycling. *arXiv preprint arXiv:2009.10748*, 2020.

[6] Pengchao Han, Shiqiang Wang, and Kin K Leung. Adaptive gradient sparsification for efficient federated learning: An online learning approach. *arXiv preprint arXiv:2001.04756*, 2020.

[7] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloe Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv:1811.03604*, 2019.

[8] Chenghao Hu, Jingyan Jiang, and Zhi Wang. Decentralized federated learning: A segmented gossip approach. *arXiv:1908.07782*, 2019.

[9] Peng Jiang and Gagan Agrawal. A linear speedup analysis of distributed deep learning with sparse and quantized communication. In *NeurIPS*, 2018.

[10] Yuang Jiang, Shiqiang Wang, Bong Jun Ko, Wei-Han Lee, and Leandros Tassiulas. Model pruning enables efficient federated learning on edge devices. *arXiv preprint arXiv:1909.12326*, 2019.

[11] Jakub Konecny, H. Brendan McMahan, Felix X. Yu, Peter Richtarik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. In *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

[12] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009.

[13] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *arXiv:1908.07873*, 2019.

[14] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks, 2020.

[15] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJxNAnVtDS

[16] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *NeurIPS*, pages 5330–5340, 2017.

[17] Xiangru Lian, Wei Zhang, Ce Zhang, and Ji Liu. Asynchronous decentralized parallel stochastic gradient descent. *arXiv preprint arXiv:1710.06952*, 2017.

[18] Lumin Liu, Jun Zhang, SH Song, and Khaled B Letaief. Client-edge-cloud hierarchical federated learning. In *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2020.

[19] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1273–1282, 2017.

[20] Swaroop Indra Ramaswamy, Rajiv Mathews, Kanishka Rao, and Francoise Beaufays. Federated learning for emoji prediction in a mobile keyboard. *arXiv:1906.04329*, 2019.
[21] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multi-task optimization under privacy constraints. arXiv:1910.01991, 2019.

[22] Jianyu Wang and Gauri Joshi. Adaptive communication strategies to achieve the best error-runtime trade-off in local-update SGD. In SysML, 2019.

[23] Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K Leung, Christian Makaya, Ting He, and Kevin Chan. Adaptive federated learning in resource constrained edge computing systems. IEEE Journal on Selected Areas in Communications, 37(6):1205–1221, 2019.

[24] Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, Proceedings of Machine Learning Research, pages 7184–7193, Long Beach, California, USA, Jun. 2019. PMLR.
Appendix

A  Proofs for the Fixed Grouping Case

We present the proofs of Theorem 1 and Corollary 1 in this section. Throughout the proof, we use the following inequalities frequently:

\[ \left\| \frac{1}{M} \sum_{i=1}^{M} x_i \right\|_2^2 \leq \frac{1}{M} \sum_{i=1}^{M} \|x_i\|^2, \quad (A.1) \]

\[ \frac{1}{M} \sum_{i=1}^{M} \|x_i - \bar{x}\|^2 = \frac{1}{M} \sum_{i=1}^{M} \|x_i\|^2 - \|\bar{x}\|^2 \leq \frac{1}{M} \sum_{i=1}^{M} \|x_i\|^2, \quad (A.2) \]

where \( \bar{x} := \frac{1}{M} \sum_{i=1}^{M} x_i \) and (A.1) is due to Jensen’s Inequality.

For the ease of notations, we define \( g_j(w) := g(w, \zeta_j) \) for any \( w \).

A.1 Proof of Theorem 1

Although the averaged global model \( \bar{w}^t \) is not observable in the system at \( t \neq aG, a = 0, 1, 2, \ldots \), here we use \( \bar{w}^t \) for analysis.

\[
\begin{align*}
\mathbb{E}f(\bar{w}^{t+1}) &= \mathbb{E}f\left(\bar{w}^t - \gamma \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j \in V_i} g_j(w_j^t)\right) \\
\quad &\leq \mathbb{E}f(\bar{w}^t) - \gamma \mathbb{E}\left(\nabla f(\bar{w}^t), \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j \in V_i} g_j(w_j^t)\right) + \frac{\gamma^2 L}{2} \mathbb{E}\left\| \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j \in V_i} g_j(w_j^t) \right\|^2, \quad (A.3)
\end{align*}
\]

where (a) is a proposition of Lipschitz smooth, which is shown in (4.3) of [3]. For the inner product term

\[
\begin{align*}
- \gamma \mathbb{E}\left(\nabla f(\bar{w}^t), \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j \in V_i} g_j(w_j^t)\right) \\
\quad &= - \gamma \mathbb{E}\left\{ \mathbb{E}\left[ \left(\nabla f(\bar{w}^t), \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j \in V_i} g_j(w_j^t)\right) \middle\| \{w_j^t : \forall j \in V\} \right] \right\} \\
\quad &= - \gamma \mathbb{E}\left(\nabla f(\bar{w}^t), \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j \in V_i} \nabla F_j(w_j^t)\right) \\
\quad &= \frac{\gamma}{2} \left( \mathbb{E}\left[ \|\nabla f(\bar{w}^t) - \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j \in V_i} \nabla F_j(w_j^t)\right|^2 - \mathbb{E}\|\nabla f(\bar{w}^t)\|^2 - \mathbb{E}\left\| \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j \in V_i} \nabla F_j(w_j^t) \right\|^2 \right). \quad (A.4)
\end{align*}
\]

Recalling that \( g_j(w_j^t) = g(w_j^t, \zeta_j^t) \), we note that the expectation operator \( \mathbb{E} \) in (A.3) is over all random samples \( \{\zeta_j^\tau : \tau = 0, 1, \cdots, t, j = 1, \cdots, n\} \), where \( \zeta_j^t \) denotes the random samples used for SGD in iteration \( t \). In step (a) of (A.4), a conditional expectation is taken for given \( \{w_j^t, j = 1, \cdots, n\} \). We note that \( \mathbb{E}\left[ g_j(w_j^t) \middle\| \{w_j^t : \forall j \in V\} \right] = \nabla F_j(w_j^t) \) due to the unbiased
gradient assumption, from which \((b)\) follows. This way of replacing the random gradient \(g_j(w^t_j)\) by its unbiased average \(\nabla F_j(w^t_j)\) through the conditional expectation is also used later in the proof, where we may not write out the conditional expectation step for compactness.

For the last term of (A.3), we have

\[
\mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j \in V_i} g_j(w^t_j) \right\|^2 \leq (a) \mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j \in V_i} (g_j(w^t_j) - \nabla F_j(w^t_j)) \right\|^2 + \mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j \in V_i} \nabla F_j(w^t_j) \right\|^2.
\]

Substitute (A.4) and (A.5) into (A.3), we have

\[
\mathbb{E} f(w^{t+1}) \leq \mathbb{E} f(w^t) + \gamma^2 \left( \frac{1}{N^2} \sum_{i=1}^{N} \frac{1}{n_i} \right) \sigma^2 - \frac{\gamma^2}{2} \mathbb{E} \|\nabla f(w^t)\|^2 - \left( \frac{\gamma}{2} - \frac{\gamma^2 L}{2} \right) \mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j \in V_i} \nabla F_j(w^t_j) \right\|^2
+ \frac{\gamma}{2} \left\| \nabla f(w^t) - \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j \in V_i} \nabla F_j(w^t_j) \right\|^2.
\]

Suppose \(\gamma \leq \frac{1}{L}\), that is, \(\frac{\gamma}{2} - \frac{\gamma^2 L}{2} \geq 0\). We can obtain

\[
\mathbb{E} f(w^{t+1}) \leq \mathbb{E} f(w^t) + \frac{\gamma^2 L}{2} \left( \frac{1}{N^2} \sum_{i=1}^{N} \frac{1}{n_i} \right) \sigma^2 - \frac{\gamma^2}{2} \mathbb{E} \|\nabla f(w^t)\|^2 + \frac{\gamma}{2} \left\| \nabla f(w^t) - \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j \in V_i} \nabla F_j(w^t_j) \right\|^2.
\]

Now we compute the upper bound of the last term of inequality (A.7)

\[
\mathbb{E} \left\| \nabla f(w^t) - \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j \in V_i} \nabla F_j(w^t_j) \right\|^2 \leq 2 \mathbb{E} \left\| \nabla f(w^t) - \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(w^t_i) \right\|^2 + 2 \mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(w^t_i) - \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j \in V_i} \nabla F_j(w^t_j) \right\|^2
\]
\[
\leq (a) 2L^2 \frac{1}{N} \sum_{l=1}^{N} \mathbb{E} \|\bar{w}^t_l - w^t_l\|^2 + 2L^2 \frac{1}{N} \sum_{l=1}^{N} \frac{1}{n_l} \sum_{k \in V_l} \mathbb{E} \|\bar{w}^t_k - w^t_k\|^2,
\]

where (a) is due to Lipschitz gradient Assumption 1(a) in the main paper and (A.1).
Substituting (A.8) to (A.7) and rearranging the order, we have

\[
\frac{\gamma}{2} \mathbb{E} \|\nabla f(\bar{w}^t)\|^2 \leq \mathbb{E} f(\bar{w}^t) - \mathbb{E} f(\bar{w}^{t+1}) + \frac{\gamma^2 L}{2} \left( \frac{1}{N^2} \sum_{i=1}^N \frac{1}{n_i} \right) \sigma^2 + \frac{\gamma}{2} \left[ 2L^2 \frac{1}{N} \sum_{l=1}^N \mathbb{E} \|\bar{w}_l - \bar{w}_l^t\|^2 \right].
\]

(A.9)

Dividing both sides by \( \frac{\gamma}{2} \) and taking the average over time, we have

\[
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\bar{w}^t)\|^2 \leq \frac{2}{\gamma} \left[ f(\bar{w}^0) - f^* \right] + L \left( \frac{1}{N^2} \sum_{i=1}^N \frac{1}{n_i} \right) \sigma^2 + 2L^2 \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{l=1}^N \mathbb{E} \|\bar{w}_l - \bar{w}_l^t\|^2 + 2L^2 \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^N \frac{1}{n_i} \sum_{k \in V_i} \mathbb{E} \|\bar{w}_k - \bar{w}_k^t\|^2,
\]

(A.10)

where \( f^* = \min_{w} f(w) \).

Note that (A.10) is the key equation that we will use to prove Theorem 1. It involves two major terms. The first is \( \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{l=1}^N \mathbb{E} \|\bar{w}_l - \bar{w}_l^t\|^2 \), which represents the global parameter mean-square-error (MSE) between the globally aggregated parameter \( \bar{w}_l \) and the locally aggregated parameters \( \{\bar{w}_l^t, l = 1, \cdots, N\} \) from \( N \) groups. The second term is \( \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^N \frac{1}{n_i} \sum_{k \in V_i} \mathbb{E} \|\bar{w}_k - \bar{w}_k^t\|^2 \), which represents the local parameter MSE between each local parameter \( \bar{w}_j \) and its locally aggregated parameter \( \bar{w}_j^t \) for each client \( j \) that belongs to group \( i \). Next, we will analyze these two terms separately.

### A.1.1 Bounding Global Parameter MSE \( \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{l=1}^N \mathbb{E} \|\bar{w}_l - \bar{w}_l^t\|^2 \) in (A.8)

In this section, we will bound \( \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{l=1}^N \mathbb{E} \|\bar{w}_l - \bar{w}_l^t\|^2 \) in terms of global divergence \( \epsilon_g^2 \) and \( \sigma^2 \). Note that when \( t = aG + c, c = 0, 1, \ldots \), we have \( \bar{w}^t = \bar{w}_j^t, l = 1, 2, \ldots, N \). Suppose \( t = aG + c, c = 0, 1, \ldots \). When \( 1 \leq c \leq G - 1 \), we have

\[
\frac{1}{N} \sum_{l=1}^N \mathbb{E} \|\bar{w}_l^t - \bar{w}_l^t\|^2 = \frac{1}{N} \sum_{l=1}^N \left\| \bar{w}_l^t - \bar{w}_l^t - \gamma \sum_{j \in V_l} \frac{1}{n_j} \sum_{m \in V_l} g_m(w_{m}^{t-1}) \right\|^2 = \frac{\gamma^2}{N} \sum_{l=1}^N \left\| \sum_{j = nG}^{nG+c-1} \left( \frac{1}{n_j} \sum_{m \in V_l} g_m(w_{m}^{t}) - \frac{1}{N} \sum_{i=1}^N \frac{1}{n_i} \sum_{j \in V_l} g_j(w_j^{t}) \right) \right\|^2.
\]

16
For the first term of (A.11), we have

\[
\frac{2\gamma^2}{N} \sum_{i=1}^{N} \left\| \sum_{\tau=\mathcal{A}G} a^{G+c-1} \left( \frac{1}{n_{li}} \sum_{m \in \mathcal{V}_i} g_m(w_m^\tau) - \nabla f_i(w_i^\tau) \right) - \left( \frac{1}{N} \sum_{i=1}^{N} \sum_{j \in \mathcal{V}_i} g_j(w_j^\tau) - \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(w_i^\tau) \right) \right\|^2.
\]

where (a) follows from (A.1) with \( M = 2 \). Next, we will derive an upper bound on each of the two terms in (A.11) separately.

For the first term of (A.11), we have

\[
\frac{2\gamma^2}{N} \sum_{i=1}^{N} \left\| \sum_{\tau=\mathcal{A}G} a^{G+c-1} \left( \frac{1}{n_{li}} \sum_{m \in \mathcal{V}_i} g_m(w_m^\tau) - \frac{1}{n_{li}} \sum_{m \in \mathcal{V}_i} \nabla F_m(w_m^\tau) + \frac{1}{n_{li}} \sum_{m \in \mathcal{V}_i} \nabla F_m(w_m^\tau) - \nabla f_i(w_i^\tau) \right) \right\|^2.
\]

For the second term of (A.11), we have

\[
\frac{4\gamma^2}{N} \sum_{i=1}^{N} \left\| \sum_{\tau=\mathcal{A}G} a^{G+c-1} \left( \frac{1}{n_{li}} \sum_{m \in \mathcal{V}_i} g_m(w_m^\tau) - \frac{1}{n_{li}} \sum_{m \in \mathcal{V}_i} \nabla F_m(w_m^\tau) \right) - \left( \frac{1}{N} \sum_{i=1}^{N} \sum_{j \in \mathcal{V}_i} g_j(w_j^\tau) - \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(w_i^\tau) \right) \right\|^2.
\]

where (a) follows from (A.1) with \( M = 2 \) and (A.2).
For the second term of (A.12), we have
\[ j \in E \]
where (A.13), (A.14), (A.15), we can obtain
\[ \leq 4\gamma^2G(1 - \frac{1}{N}) \sum_{i=1}^{N} \frac{1}{n_i} \sigma^2 \] (use Assumption 1b),  
\[ (A.13) \]

where (a) is due to (A.2), (b) and (c) hold due to the independence of SGD noise over different time, groups and nodes, and \( \mathbb{E}[g_i(w^t_i)] = \mathbb{E}[\mathbb{E}[g_i(w^t_i)|w^t_i]] = \mathbb{E}[\nabla F_i(w^t_i)] \), \( \forall i, t \). This implies that \( \mathbb{E}[g_i(w^t_i) - \nabla F_i(w^t_i), g_i(w^t'_i) - \nabla F_i(w^t'_i)] = 0, i \neq j \).

For the second term of (A.12), we have
\[
\frac{4\gamma^2}{N} \sum_{i=1}^{N} \mathbb{E} \left[ \left( \frac{1}{n_i} \sum_{m \in \mathcal{V}_i} \nabla F_m(w^t_m) - \nabla f_i(w^t_i) \right)^2 \right] \\
\leq \frac{4\gamma^2 G L^2}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{m \in \mathcal{V}_i} \mathbb{E}[\|\nabla F_m(w^t_m) - \nabla F_m(w^t_i)\|^2] \\
\text{(use (A.1) with } M = G \text{)} \\
\leq \frac{4\gamma^2 G L^2}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{m \in \mathcal{V}_i} \mathbb{E}[\|w^t_m - w^t_i\|^2] \\
\text{(use (A.1) with } M = n_i \text{ and Assumption 1a)}.
\[ (A.14) \]

For the second term of (A.11), we have
\[
\frac{2\gamma^2}{N} \sum_{i=1}^{N} \mathbb{E} \left[ \left( \frac{1}{n_i} \sum_{m \in \mathcal{V}_i} \nabla F_m(w^t_m) - \nabla f_i(w^t_i) \right)^2 \right] \\
\leq \frac{2\gamma^2}{N} \sum_{i=1}^{N} \mathbb{E} \left[ \left( \frac{1}{n_i} \sum_{m \in \mathcal{V}_i} \nabla F_m(w^t_m) - \nabla f_i(w^t_i) \right)^2 \right] \\
\leq 6\gamma^2G^2\epsilon^2 + \frac{6\gamma^2L^2G}{N} \sum_{\tau = aG}^{aG+c-1} \frac{1}{n_i} \sum_{m \in \mathcal{V}_i} \mathbb{E}[\|\bar{w}^\tau - \bar{w}^t\|^2] \\
= 6\gamma^2G^2\epsilon^2 + \frac{12\gamma^2L^2G}{N} \sum_{\tau = aG}^{aG+c-1} \sum_{i=1}^{N} \mathbb{E}[\|\bar{w}^\tau - \bar{w}^t\|^2],
\[ (A.15) \]

where (a) uses (A.1) with \( M = 3 \) and \( M = G \), and Assumptions 1b and 1a.

Now back to inequalities (A.11), with results of inequality (A.12), (A.13), (A.14), (A.15), we can obtain
\[
\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[\|w^t_i - \bar{w}^t_i\|^2]
\]
\[ 18 \]
where \( a \leq c \leq G - 1 \). Since \( \| \tilde{w}^t - \bar{w}^t \|_2 = 0 \), \((a)\) also holds when \( t = aG, a = 0, 1, 2, \ldots \). Suppose \( T - 1 = dG, d = 0, 1, 2, \ldots \). Take the average over time on both sides of (A.16), we have

\[
\frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\|w^t - \bar{w}^t\|^2 = \frac{1}{T} \sum_{t=0}^{T-2} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\|w^t - \bar{w}^t\|^2
\]

\[(a)\] \[
\leq 4\gamma^2 G \left( 1 - \frac{1}{N} \right) \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sigma^2 + 6\gamma^2 G \epsilon^2 \left( 1 + \frac{4\gamma^2 G^2 \epsilon}{\sqrt{N}} \right) \sum_{i=1}^{N} \frac{1}{n_i} \sum_{m \in V_i} \mathbb{E}\|w^t_m - \bar{w}^t\|^2
\]

\[
+ \frac{12\gamma^2 L^2 G \epsilon}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{m \in V_i} \mathbb{E}\|w^t_m - \bar{w}^t\|^2
\]

\[
\leq 4\gamma^2 G \left( 1 - \frac{1}{N} \right) \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sigma^2 + 6\gamma^2 G \epsilon^2 \left( 1 + \frac{4\gamma^2 G^2 \epsilon}{\sqrt{N}} \right) \sum_{i=1}^{N} \frac{1}{n_i} \sum_{m \in V_i} \mathbb{E}\|w^t_m - \bar{w}^t\|^2
\]

\[
+ \frac{12\gamma^2 L^2 G \epsilon}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{m \in V_i} \mathbb{E}\|w^t_m - \bar{w}^t\|^2
\]

\[(b)\] \[
= 4\gamma^2 G \left( 1 - \frac{1}{N} \right) \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sigma^2 + 6\gamma^2 G \epsilon^2 \left( 1 + \frac{4\gamma^2 G^2 \epsilon}{\sqrt{N}} \right) \sum_{i=1}^{N} \frac{1}{n_i} \sum_{m \in V_i} \mathbb{E}\|w^t_m - \bar{w}^t\|^2
\]

\[
+ \frac{12\gamma^2 L^2 G \epsilon}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{m \in V_i} \mathbb{E}\|w^t_m - \bar{w}^t\|^2
\]

\[(A.17)\]

where \((a)\) and \((b)\) hold due to \( \|w^t_m - \bar{w}^t\|_2 = 0 \) and \( \|\bar{w}^t_m - \bar{w}^t\|_2 = 0 \). Suppose \( 1 - 12\gamma^2 L^2 G^2 > 0 \), move the last term of (A.17) to the left, we can obtain

\[
(1 - 12\gamma^2 L^2 G^2) \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\|w^t - \bar{w}^t\|^2
\]

\[
\leq 4\gamma^2 G \left( 1 - \frac{1}{N} \right) \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sigma^2 + 6\gamma^2 G \epsilon^2 \left( 1 + \frac{4\gamma^2 G^2 \epsilon}{\sqrt{N}} \right) \sum_{i=1}^{N} \frac{1}{n_i} \sum_{m \in V_i} \mathbb{E}\|w^t_m - \bar{w}^t\|^2,
\]

(A.18)
then

\[
\frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \| \tilde{w}_t - \bar{w}_t \|^2 \\
\leq \frac{4\gamma^2 G(1 - \frac{1}{N}) \sum_{i=1}^{N} \frac{1}{n_i} \sigma^2}{1 - 12\gamma^2 L^2 G^2} + \frac{6\gamma^2 G^2}{1 - 12\gamma^2 L^2 G^2} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{m \epsilon V_i} \mathbb{E} \| w_m^t - \bar{w}_t \|^2.
\]

(A.19)

The importance of (A.19) is that it presents an upper bound on the global parameter MSE in terms of the local parameter MSE. Next, we will proceed to analyze the local parameter MSE.

A.1.2 Bounding Local parameter MSE

\[\frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{n_i} \sum_{k \epsilon V_i} \mathbb{E} \| w_t^i - w_t^i \|^2 \text{ in (A.8)} \]

Now we consider the local aggregation. Suppose \( t = b_i I_i + c_i \) for \( i = 1, 2, 3, \ldots, N \) where \( b_i = 0, 1, \ldots, \frac{N}{I_i} - 1 \) and \( c_i = 0, 1, 2, \ldots, I_i - 1 \). When \( c_i = 0 \), we have \( w_t^i = w_t^k, k \epsilon V_i, \forall i \). When \( c_i \geq 1 \), we have

\[
\frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \epsilon V_i} \mathbb{E} \| w_t^i - w_t^i \|^2
\]

\[
= \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \epsilon V_i} \mathbb{E} \left\| (w_t^{i-1} - \gamma \frac{1}{n_i} \sum_{j \epsilon V_i} g_j(w_t^{i-1})) - (w_t^{k-1} - \gamma g_k(w_t^{k-1})) \right\|^2
\]

\[
= \frac{\gamma^2}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \epsilon V_i} \mathbb{E} \left\| \sum_{\tau=b_i I_i}^{t-1} \left( g_k(w_t^\tau) - \frac{1}{n_i} \sum_{j \epsilon V_i} g_j(w_t^\tau) \right) \right\|^2
\]

\[
= \frac{\gamma^2}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \epsilon V_i} \mathbb{E} \left\| \sum_{\tau=b_i I_i}^{t-1} \left( g_k(w_t^\tau) - \nabla F_k(w_t^\tau) + \nabla F_k(w_t^\tau) - \frac{1}{n_i} \sum_{j \epsilon V_i} \nabla F_j(w_t^\tau) + \frac{1}{n_i} \sum_{j \epsilon V_i} \nabla F_j(w_t^\tau) - \frac{1}{n_i} \sum_{j \epsilon V_i} g_j(w_t^\tau) \right) \right\|^2
\]

\[
= \frac{\gamma^2}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \epsilon V_i} \mathbb{E} \left\| \sum_{\tau=b_i I_i}^{t-1} \left( g_k(w_t^\tau) - \nabla F_k(w_t^\tau) + \frac{1}{n_i} \sum_{j \epsilon V_i} \nabla F_j(w_t^\tau) - \frac{1}{n_i} \sum_{j \epsilon V_i} g_j(w_t^\tau) + \nabla F_k(w_t^\tau) - \frac{1}{n_i} \sum_{j \epsilon V_i} \nabla F_j(w_t^\tau) \right) \right\|^2
\]

\[
\leq \frac{\gamma^2}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \epsilon V_i} \mathbb{E} \left\| \sum_{\tau=b_i I_i}^{t-1} \left( g_k(w_t^\tau) - \nabla F_k(w_t^\tau) + g_j(w_t^\tau) - \nabla F_j(w_t^\tau) \right) \right\|^2
\]

(A.20)

\[
\leq \frac{2\gamma^2}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \epsilon V_i} \mathbb{E} \left\| \sum_{\tau=b_i I_i}^{t-1} \left( \nabla F_k(w_t^\tau) - \frac{1}{n_i} \sum_{j \epsilon V_i} \nabla F_j(w_t^\tau) \right) \right\|^2
\]

(A.21)

For (A.20), we have

\[
\frac{2\gamma^2}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \epsilon V_i} \mathbb{E} \left\| \sum_{\tau=b_i I_i}^{t-1} \left( \nabla F_k(w_t^\tau) - \frac{1}{n_i} \sum_{j \epsilon V_i} \nabla F_j(w_t^\tau) \right) \right\|^2
\]

20
For (A.21), we have

\[
\frac{2\gamma^2}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \in V_i} E \left\| \sum_{\tau=b_i I_i}^{t-1} \left( g_k(w_k^\tau) - \nabla F_k(w_k^\tau) \right) \right\|^2 - \frac{2\gamma^2}{N} \sum_{i=1}^{N} E \left\| \sum_{\tau=b_i I_i}^{t-1} \left( \frac{1}{n_i} \sum_{j \in V_i} g_j(w_j^\tau) - \frac{1}{n_i} \sum_{j \in V_i} \nabla F_j(w_j^\tau) \right) \right\|^2
\]

\[
= \frac{2\gamma^2}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \in V_i} \sum_{\tau=b_i I_i}^{t-1} E \| g_k(w_k^\tau) - \nabla F_k(w_k^\tau) \|^2 - \frac{2\gamma^2}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j \in V_i} \sum_{\tau=b_i I_i}^{t-1} E \| g_j(w_j^\tau) - \nabla F_j(w_j^\tau) \|^2
\]

\[
= \frac{2\gamma^2}{N} \sum_{i=1}^{N} \left( \frac{1}{n_i} - \frac{1}{n_i^2} \right) \sum_{k \in V_i} \sum_{\tau=b_i I_i}^{t-1} E \| g_k(w_k^\tau) - \nabla F_k(w_k^\tau) \|^2
\]

\[
\leq \frac{2\gamma^2}{N} \sum_{i=1}^{N} \left( 1 - \frac{1}{n_i} \right) I_i \sigma^2. \tag{A.22}
\]

For (A.21), we have

\[
\frac{2\gamma^2}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \in V_i} E \left\| \sum_{\tau=b_i I_i}^{t-1} \left( \nabla F_k(w_k^\tau) - \frac{1}{n_i} \sum_{j \in V_i} \nabla F_j(w_j^\tau) \right) \right\|^2
\]

\[
= \frac{2\gamma^2}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \in V_i} \sum_{\tau=b_i I_i}^{t-1} E \left\| \sum_{\tau=b_i I_i}^{t-1} \left( \nabla F_k(w_k^\tau) - \nabla F_k(w_k^\tau) + \nabla F_k(w_k^\tau) - \nabla f_i(w_i^\tau) + \nabla f_i(w_i^\tau) - \frac{1}{n_i} \sum_{j \in V_i} \nabla F_j(w_j^\tau) \right) \right\|^2
\]

\[
\leq \frac{6\gamma^2}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \in V_i} \sum_{\tau=b_i I_i}^{t-1} E \left\| \nabla f_i(w_i^\tau) - \frac{1}{n_i} \sum_{j \in V_i} \nabla F_j(w_j^\tau) \right\|^2 + \frac{6\gamma^2}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \in V_i} \sum_{\tau=b_i I_i}^{t-1} E \left\| \sum_{\tau=b_i I_i}^{t-1} \left( \nabla F_k(w_k^\tau) - \nabla F_k(w_k^\tau) \right) \right\|^2
\]

\[
\leq \frac{12\gamma^2}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \in V_i} \sum_{\tau=b_i I_i}^{t-1} I_i \| w_k^\tau - w_i^\tau \|^2 + \frac{6\gamma^2}{N} \sum_{i=1}^{N} I_i^2 \tau_i^2 \tag{A.23}
\]

where (a) is due to (A.1) with \( M = I_i \) and \( M = n_i \) respectively.

Substituting (A.22) and (A.23) back to (A.20) and (A.21), we have

\[
\frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \in V_i} E \| \bar{w}_i^\tau - w_i^\tau \|^2
\]

\[
\leq \frac{2\gamma^2}{N} \sum_{i=1}^{N} \left( 1 - \frac{1}{n_i} \right) I_i \sigma^2 + \frac{6\gamma^2}{N} \sum_{i=1}^{N} I_i^2 \tau_i^2 + \frac{12\gamma^2}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \in V_i} \sum_{\tau=b_i I_i}^{t-1} I_i E \| w_k^\tau - w_i^\tau \|^2
\]

\[
\leq \frac{2\gamma^2}{N} \sum_{i=1}^{N} \left( 1 - \frac{1}{n_i} \right) I_i \sigma^2 + \frac{6\gamma^2}{N} \sum_{i=1}^{N} I_i^2 \tau_i^2 + \frac{12\gamma^2}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \in V_i} \sum_{\tau=b_i I_i}^{t-1} I_i E \| w_k^\tau - w_i^\tau \|^2 \tag{A.24}
\]
where (a) holds due to $c_i \leq I_i - 1$. (A.24) also holds when $t = b_i I_i$, since $\frac{1}{n_i} \sum_{k \in V_i} E \|w_i^{b_i I_i} - w_k^{b_i I_i}\|^2 = 0$.

Since $T - 1 = dG, d = 0, 1, \ldots, T - 1$ is also an integral multiple of $I_i$. That is, $T - 1 = f_i I_i$. Taking the average over time yields

$$
\frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \in V_i} E \|w_i^t - w_k^t\|^2 \\
\leq \frac{2\gamma^2}{N} \sum_{i=1}^{N} \left(1 - \frac{1}{n_i}\right) I_i \sigma^2 + \frac{6\gamma^2}{N} \sum_{i=1}^{N} I_i^2 \epsilon_i^2 + \frac{12\gamma^2 L^2}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \in V_i} \sum_{\tau = b_i I_	au}^{f_i - 1} I_i^2 \|w_k^\tau - w_i^\tau\|^2 \\
\leq \frac{2\gamma^2}{N} \sum_{i=1}^{N} \left(1 - \frac{1}{n_i}\right) I_i \sigma^2 + \frac{6\gamma^2}{N} \sum_{i=1}^{N} I_i^2 \epsilon_i^2 + \frac{1}{T} \sum_{\tau = 0}^{T-1} \frac{12\gamma^2 L^2}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \in V_i} I_i^2 \|w_k^\tau - w_i^\tau\|^2 \\
\leq \frac{2\gamma^2}{N} \sum_{i=1}^{N} \left(1 - \frac{1}{n_i}\right) I_i \sigma^2 + \frac{6\gamma^2}{N} \sum_{i=1}^{N} I_i^2 \epsilon_i^2 + \frac{1}{T} \sum_{\tau = 0}^{T-1} \frac{12\gamma^2 L^2 I_{max}}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \in V_i} \|w_k^\tau - w_i^\tau\|^2, \quad (A.25)
$$

where $I_{max} = \max_i I_i$. Suppose $1 - 12\gamma^2 L^2 I_{max}^2 > 0$, we have

$$
\frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{k \in V_i} E \|w_i^t - w_k^t\|^2 \\
\leq \frac{2\gamma^2}{N} \sum_{i=1}^{N} \left(1 - \frac{1}{n_i}\right) I_i \sigma^2 + \frac{6\gamma^2}{N} \sum_{i=1}^{N} I_i^2 \epsilon_i^2 + \frac{6\gamma^2 \frac{1}{N} \sum_{i=1}^{N} I_i^2 \epsilon_i^2}{1 - 12\gamma^2 L^2 I_{max}^2}. \quad (A.26)
$$

### A.1.3 Obtaining the Final Result

Substituting (A.26) into (A.19), we obtain an upper bound on the global parameter MSE as

$$
\frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^{N} E \|\hat{w}^t - w_i^t\|^2 \\
\leq \frac{4\gamma^2 G (1 - \frac{1}{N})}{1 - 12\gamma^2 L^2 G^2} \sum_{i=1}^{N} \frac{1}{n_i} \sigma^2 + \frac{6\gamma^2 G^2 \epsilon_i^2}{1 - 12\gamma^2 L^2 G^2} \\
+ \frac{4\gamma^2 G^2 L^2}{1 - 12\gamma^2 L^2 G^2} \left(2\frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} I_i \sigma^2 + \frac{6\gamma^2 \frac{1}{N} \sum_{i=1}^{N} I_i^2 \epsilon_i^2}{1 - 12\gamma^2 L^2 I_{max}^2}\right). \quad (A.27)
$$

Taking the average over time and substitute (A.26) and (A.27) to (A.10), we obtain (9a)--(9c) of the main paper:

$$
\frac{1}{T} \sum_{t=0}^{T-1} E \|\nabla f(\hat{w}^t)\|^2 \\
\leq \frac{2\gamma}{N} (E f(w^0) - f^*) + \gamma L \frac{1}{N^2} \sum_{i=1}^{N} \frac{1}{n_i} \sigma^2
$$
\[
\frac{8\gamma^2 L^2 G (1 - \frac{1}{N})}{1 - 12\gamma^2 L^2 G^2} \sum_{i=1}^{N} \frac{1}{n_i} \sigma_i^2 + \frac{12\gamma^2 L^2 G^2 \epsilon_i^2}{1 - 12\gamma^2 L^2 G^2} + \left(1 + \frac{8\gamma^2 G^2 L^2}{1 - 12\gamma^2 L^2 G^2} \right) \left( \frac{4\gamma^2 L^2 \sum_{i=1}^{N} \left(1 - \frac{1}{n_i}\right) I_i \sigma_i^2}{1 - 12\gamma^2 L^2 I_{\text{max}}^2} + \frac{12\gamma^2 L^2 \sum_{i=1}^{N} I_i^2 \epsilon_i^2}{1 - 12\gamma^2 L^2 I_{\text{max}}^2} \right).
\] (A.28)

This completes the proof of Theorem 1.

A.2 Proof of Corollary 1

When \(0 < \gamma \leq \frac{1}{2\sqrt{6}GL}\), we have

\[
0 < \frac{8\gamma^2 G^2 L^2}{1 - 12\gamma^2 L^2 G^2} \leq \frac{2}{3},
\] (A.29)

and

\[
0 < \frac{1}{1 - 12\gamma^2 L^2 I_{\text{max}}^2} \leq \frac{1}{1 - 12\gamma^2 L^2 G^2} \leq 2.
\] (A.30)

By substituting (A.29) and (A.30) into (9a)–(9c) of the main paper and setting \(C = \frac{40}{3}\), we obtain an upper bound in the form of (10) in Corollary 1. Note that the constant in front of the term \(\frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{1}{n_i}\right) I_i\) is intentionally increased by a factor of 2 in order to better match the constants given in Corollary 2.
B Proofs for the Random Grouping Case

In this section, we give the proofs of Lemmas 1 and 2 and Theorem 2. As discussed in Section 4.3 of the main paper, we consider uniformly random grouping here where the groups \{V_i : i = 1, 2, \ldots, N\} are random sets of equal size, i.e., \(n_i = |V_i| = \frac{N}{K}\) for all \(i\). For simplicity, we use \(K := \frac{N}{K} = n_i\) to denote the size of each group. The uniformly random partitioning of clients into groups is equivalent to the following problem. Assume that there are \(n\) bins, where the first \(K\) bins contain clients in group \(i = 1\), the next \(K\) bins contain clients in group \(i = 2\), and so on. We randomly and uniformly assign the \(n\) clients to the \(n\) bins, such that each bin contains one (and only one) client and each client is assigned to one (and only one) bin. A random realization \(S\) corresponds to a mapping between clients and bins.

Let \(k_j\) denote the client assigned to bin \(j\), where \(k_j\) is a random variable due to random assignment. Also let \([K] := \{1, 2, ..., K\}\) and \([n] := \{1, 2, ..., n\} = \mathcal{V}\) (the last equality is by definition of \(\mathcal{V}\)). We proceed with the proofs in the following.

We define

\[
\epsilon_w^2 := \mathbb{E}_S \left( \| \nabla f(w) - \nabla F_{k_j}(w) \|\right)^2 \leq \frac{1}{n} \sum_{k=1}^{n} \| \nabla f(w) - \nabla F_{k_j}(w) \|^2 \leq \epsilon^2.
\]

where (a) is because for any given bin \(j\), the probability of any client being assigned to it is \(\frac{1}{n}\), and (b) is due to Assumption 2 in the main paper.

B.1 Proof of Lemma 1

We have

\[
\begin{align*}
\mathbb{E}_S \left[ \frac{1}{N} \sum_{i=1}^{N} \left\| \nabla f(w) - \frac{1}{K} \sum_{k \in V_i} \nabla F_k(w) \right\|^2 \right] \\
= \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_S \left\| \nabla f(w) - \frac{1}{K} \sum_{k \in V_i} \nabla F_k(w) \right\|^2 \\
\overset{(a)}{=} \mathbb{E}_S \left\| \nabla f(w) - \frac{1}{K} \sum_{k \in V_i} \nabla F_k(w) \right\|^2 \\
\overset{(b)}{=} \mathbb{E}_S \left\| \nabla f(w) - \frac{1}{K} \sum_{k \in V_i} \nabla F_k(w) \right\|^2 \\
= \frac{1}{K^2} \mathbb{E}_S \left[ \sum_{j=1}^{K} (\nabla f(w) - \nabla F_{k_j}(w)) \right]^2 \\
= \frac{1}{K^2} \mathbb{E}_S \left( \sum_{j=1}^{K} \| \nabla f(w) - \nabla F_{k_j}(w) \|^2 \right) + \frac{1}{K^2} \mathbb{E}_S \left( \sum_{j, j' \in [K], j \neq j'} \langle \nabla f(w) - \nabla F_{k_j}(w), \nabla f(w) - \nabla F_{k_j'}(w) \rangle \right),
\end{align*}
\]

where we note that \(V_i\) is a random set of clients in group \(i\) and \(k_j\) is a random client index assigned to the \(j\)-th bin, both of which capture the outcome of random grouping. Equality (a) is due to symmetry, so that the expectation involving any group.
i is equal to the expectation involving the first group \( i = 1 \). Equality \((b)\) is from the fact that the first group \( \mathcal{V}_1 \) includes clients in bins \( j = 1, 2, \ldots, K \), hence \( \sum_{k \in \mathcal{V}_1} \nabla F_k(w) = \sum_{j=1}^{K} \nabla F_{kj}(w) \). We analyze the two terms in (B.2) in the following.

For the first term in (B.2), we have

\[
\frac{1}{K^2} \mathbb{E}_S \left( \sum_{j=1}^{K} \||\nabla f(w) - \nabla F_{kj}(w)||^2 \right) = \frac{1}{K^2} \sum_{j=1}^{K} \mathbb{E}_S \left( ||\nabla f(w) - \nabla F_{kj}(w)||^2 \right) = \frac{c^2_w}{K}, \tag{B.3}
\]

where the last inequality is from (B.1).

For the second term in (B.2), we have

\[
\frac{1}{K^2} \mathbb{E}_S \left( \sum_{j,j' \in [K], j \neq j'} \left< \nabla f(w) - \nabla F_{kj}(w), \nabla f(w) - \nabla F_{k'j'}(w) \right> \right)
= \frac{1}{K^2} \sum_{j,j' \in [K], j \neq j'} \mathbb{E}_S \left< \nabla f(w) - \nabla F_{kj}(w), \nabla f(w) - \nabla F_{k'j'}(w) \right>
= \frac{1}{K^2} \sum_{j,j' \in [K], j \neq j'} \frac{1}{n(n-1)} \left| \sum_{k=1}^{n} \left< \nabla f(w) - \nabla F_k(w), \nabla f(w) - \nabla F_k(w) \right> \right|^2
\]

\[
= \frac{1}{K^2} \sum_{j,j' \in [K], j \neq j'} \frac{1}{n(n-1)} \left| \sum_{k=1}^{n} \nabla f(w) - \nabla F_k(w) \right|^2
= - \frac{K-1}{K} \cdot \frac{1}{n-1} \cdot \frac{1}{n} \sum_{k=1}^{n} \nabla f(w) - \nabla F_k(w) \right|^2
\]

\[
= - \frac{K-1}{K} \cdot \frac{1}{n-1} c_w^2, \tag{B.4}
\]

where \((a)\) is because the probability of assigning client \( k \) to bin \( j \) and at the same time a different client \( k' \) to a different bin \( j' \) is \( \frac{1}{n(n-1)} \), \((b)\) is because \( \sum_{k=1}^{n} \left< \nabla f(w) - \nabla F_k(w) \right> = 0 \) according to the definition in (1) and the fact that \( n_i \) is equal for all \( i \), and \((c)\) is from (B.1).

Substituting (B.3) and (B.4) into (B.2), we obtain

\[
\mathbb{E}_S \left[ \frac{1}{N} \sum_{i=1}^{N} \left< \nabla f(w) - \frac{1}{K} \sum_{k \in \mathcal{V}_i} \nabla F_k(w) \right> \right]^2
= \frac{1}{K} \left[ c^2_w - \frac{K-1}{n-1} c^2_w \right]
\]

25
\[
\begin{align*}
&= \frac{1}{K} \left( \frac{n-K}{n-1} \right) \epsilon_w^2 = \left( \frac{N-1}{n-1} \right) \epsilon_w^2 \leq \left( \frac{N-1}{n-1} \right) \epsilon^2, \\
\end{align*}
\]
where \((a)\) is from (B.1). This proves the lemma.

### B.2 Proof of Lemma 2

We have

\[
\begin{align*}
\mathbb{E}_S \left[ \frac{1}{K} \sum_{k \in V_i} \left\| \nabla F_k(w) - \frac{1}{K} \sum_{k' \in V_i} \nabla F_{k'}(w) \right\|^2 \right] \\
&= \frac{1}{K} \sum_{k \in V_i} \mathbb{E}_S \left[ \left\| \nabla F_k(w) - \frac{1}{K} \sum_{k' \in V_i} \nabla F_{k'}(w) \right\|^2 \right] \\
&= \frac{1}{K} \sum_{j=1}^{K} \sum_{k \in V_i} \mathbb{E}_S \left[ \left\| \nabla F_{k_j}(w) - \frac{1}{K} \sum_{j' \neq j} \nabla F_{k_{j'}}(w) \right\|^2 \right] \\
&= \frac{1}{K} \sum_{j=1}^{K} \mathbb{E}_S \left[ \left\| \nabla F_{k_j}(w) - \nabla f(w) + \nabla f(w) - \frac{1}{K} \sum_{j' \neq j} \nabla F_{k_{j'}}(w) \right\|^2 \right] \\
&= \frac{1}{K} \sum_{j=1}^{K} \mathbb{E}_S \left[ \left\| \nabla F_{k_j}(w) - \nabla f(w) \right\|^2 \right] + \mathbb{E}_S \left[ \left\| \nabla f(w) - \frac{1}{K} \sum_{j' \neq j} \nabla F_{k_{j'}}(w) \right\|^2 \right] \\
&\quad + 2 \mathbb{E}_S \left\langle \frac{1}{K} \sum_{j=1}^{K} \nabla F_{k_j}(w) - \nabla f(w), \nabla f(w) - \frac{1}{K} \sum_{j' \neq j} \nabla F_{k_{j'}}(w) \right\rangle \\
&= \frac{1}{K} \sum_{j=1}^{K} \mathbb{E}_S \left[ \left\| \nabla F_{k_j}(w) - \nabla f(w) \right\|^2 \right] + \mathbb{E}_S \left[ \left\| \nabla f(w) - \frac{1}{K} \sum_{j' \neq j} \nabla F_{k_{j'}}(w) \right\|^2 \right] \\
&\quad + 2 \mathbb{E}_S \left\langle \frac{1}{K} \sum_{j=1}^{K} \nabla F_{k_j}(w) - \nabla f(w), \nabla f(w) - \frac{1}{K} \sum_{j' \neq j} \nabla F_{k_{j'}}(w) \right\rangle \\
&\leq \left( 1 - \frac{N-1}{n-1} \right) \epsilon_w^2, \\
\end{align*}
\]

where \((a)\) is due to symmetry of groups so that the expectation over \(V_i\) is equal to the expectation over \(V_1\), \((b)\) is because the clients in group \(i = 1\) are in the first \(K\) bins, the first term in \((c)\) is from (B.1), the second term in \((c)\) is from the intermediate result in (B.5) of Lemma 1’s proof and also the symmetry of groups, and \((d)\) is from (B.1) and \(1 - \frac{N-1}{n-1} \geq 0.\)
B.3 Proof of Theorem 2

From Lemma 1 and Lemma 2, we have the expected global divergence and local divergence:

\[
\mathbb{E}_S \left[ \frac{1}{N} \sum_{i=1}^{N} \| \nabla f_i(w) - \nabla f(w) \|_2^2 \right] \leq \left( \frac{N - 1}{n - 1} \right) \tilde{\epsilon}^2, \forall w, \tag{B.7}
\]

\[
\mathbb{E}_S \left[ \frac{1}{n_i} \sum_{k \in V_i} \| \nabla F_k(w) - \nabla f_i(w) \|_2^2 \right] \leq \left( 1 - \frac{N - 1}{n - 1} \right) \tilde{\epsilon}^2, \forall w, \forall i, \tag{B.8}
\]

where we recall that \( n_i = K = \frac{n}{N} \) and \( f_i(w) = \frac{1}{n_i} \sum_{j \in V_i} F_j(w) \) by definition.

The proof of Theorem 2 follows similar steps as that of Theorem 1 and Corollary 1, with the following differences: 1) we take an additional expectation over \( S \) (i.e., all possible groupings) and consider the expectations currently in Theorem 1, Corollary 1, and their proofs as being conditioned on a given grouping realization \( S \); 2) we substitute (B.7) and (B.8) into (A.15) and (A.23) to replace the original upper bounds \( \epsilon^2 \) and \( \epsilon_i^2 \) with the expected divergence upper bounds \( \left( \frac{N - 1}{n - 1} \right) \tilde{\epsilon}^2 \) and \( \left( 1 - \frac{N - 1}{n - 1} \right) \tilde{\epsilon}^2 \), respectively. Due to the similarity, we do not write the full proof of Theorem 2 and refer readers to the proofs of Theorem 1 and Corollary 1 while noting the above minor changes.
C  Additional Experiments with Partial Client Participation

In this section, we provide experiments with partial client participation in FL with random client sampling in each round, as shown in Figures C.1–C.3. We separate data into 50 clients in non-IID manner. There are the same number of clients in each group. For each FL round, we uniformly sample $20\%$ of clients in each group. The results show that the same insights as described in Section 6.2 of the main paper can be observed here.

![Figure C.1: Comparison between SF-SGD and HF-SGD.](image1)

![Figure C.2: Comparison between SF-SGD and HF-SGD with a large $G$.](image2)
Figure C.3: Comparison between different number of groups.