Hierarchical Federated Learning: The Interplay of User Mobility and Data Heterogeneity

Wei Dong and Howard H. Yang
ZJU-UIUC Institute, Zhejiang University
Email: {weid.21, haoyang}@intl.zju.edu.cn

Chenyuan Feng
EURECOM, France
Email: Chenyuan.Feng@eurecom.fr

Chen Sun
Beijing Lab, Sony
Email: chen.sun@sony.com

***Abstract*—THIS PAPER IS ELIGIBLE FOR THE STUDENT PAPER AWARD. Federated Learning (FL) is envisioned as the cornerstone of the next-generation mobile system, whereby integrating FL into the network edge elements (i.e., user terminals and edge/cloud servers), it is expected to unleash the potential of network intelligence by learning from the massive amount of users' data while concurrently preserving privacy. In this paper, we develop an analytical framework that quantifies the interplay of user mobility, a fundamental property of mobile networks, and data heterogeneity, the salient feature of FL, on the model training efficiency. Specifically, we derive the convergence rate of a hierarchical FL system operated in a mobile network, showing how user mobility amplifies the divergence caused by data heterogeneity. The theoretical findings are corroborated by experimental simulations.**

I. Introduction

Federated learning (FL) is an emerging paradigm of distributed machine learning that enables a large number of end users to collaboratively learn a global model without directly accessing their raw data [1]–[4]. Inspired by FL, there is a trend of upgrading the wireless network by integrating FL into the mobile system, achieving network intelligence. Under such a context, mobile devices shall execute local model updates and upload the intermediate parameters to edge servers, preserving user privacy while reducing communication strain on the network as well as alleviating computational pressure on the server side. To extend FL into a large-scale deployment with massive users, hierarchy federated learning (HFL) is introduced to fully exploit the training data of mobile devices [5]–[7]. In wireless HFL, mobile users directly communicate with the nearby edge server rather than the remote cloud server, which results in a higher communication efficiency with more involved users.

By nature, users roam in a mobile network; this mobility inevitably decreases the number of users participating in the model training process of each edge server, thereby degrading the FL performance [8]–[11]. Aside from user mobility, HFL also confronts the challenge of data heterogeneity, which arises from end users' personal preferences and leads to non-identically and independently distributed (non-IID) data sets. Albeit the model training of FL might still converge under non-IID data sets, a significant degradation of model accuracy occurs as local data distributions diverge [12]–[15]. In the presence of user mobility, performance degradation in non-IID scenarios becomes more evident, whereas the degradation further exacerbates when users have higher mobility. This phenomenon has been empirically observed in [8], without a theoretical explanation.

The present paper closes this research gap by developing a theoretical framework that provides an acute understanding of the interplay of user mobility and data heterogeneity on HFL. Specifically, we show that system divergence of HFL mainly comes from two aspects: local updates due to data heterogeneity and connectivity loss due to user mobility. Our analysis also reveals that the connectivity loss incurred by user mobility is associated with the *downward divergence*, a notion introduced in [15], characterizing the data heterogeneity within an edge server. The connectivity loss will be amplified by the local divergence introduced by data heterogeneity, which explains why higher user mobility dramatically drops the algorithm performance in a non-IID scenario, as reported in [8]. In brief, the principal contributions of this paper are summarized below:

1) We establish a theoretical framework for analyzing the training efficiency of an HFL system, encompassing critical factors such as user mobility, data heterogeneity, and system hierarchy.
2) We derive the convergence rate of model training under the considered HFL system. The analysis characterizes the interplay between user mobility and data heterogeneity, showing how the connectivity loss enlarge the influence of data heterogeneity.

A. Related Works

This part conducts a brief survey on the current progress of investigating the effects of non-IID data and user mobility on FL performance.

1) Federated Learning with Non-IID Data: For single-layer FL, model errors are compared between centralized and FL models in [13], revealing performance degradation with increased data distribution distance. This distance serves as a metric for non-IID degrees. Convergence rates of FedAvg with non-IID data are analyzed in [12]. In the context of HFL, effects of non-IID are analyzed in [5], in which the notions of upward and downward divergences have been introduced to assess the HFL data heterogeneity.

2) Federated Learning with Mobile Users: Mobility induces dynamics in network topology, causing connectivity loss and inconsistent collaboration [9]. The degradation in

link quality leads to more frequent transmission failures and deteriorating system performance [10]. To overcome this challenge, a reputation-based worker selection scheme is proposed in [10]. Additionally, user selection and wireless resource allocation are optimized based on user locations during FL model transmission [16]. A cluster FL algorithm is proposed in [8], involving collaboration among multiple edge servers in FL training. Notably, [11] points out slight user mobility can positively impact FL, but this analysis considers only one base station.

II. System Model

A. Network Setup

We consider the HFL system depicted in Fig. 1, consisting of one cloud server, M edge servers, and N mobile users. Each user holds a local dataset $\mathcal{D}_n$ with size $|\mathcal{D}_n|$. And the users aim to collaboratively train a global model $\boldsymbol{w}$ by minimizing the following global loss functions, $f(\boldsymbol{w}, \mathcal{D})$ while preserving their data privacy:

$$f(\boldsymbol{w}, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x}_i, y_i) \in \mathcal{D}} \ell(\boldsymbol{w}; \boldsymbol{x}_i, y_i) = \sum_{n=1}^{N} \alpha_n F_n(\boldsymbol{w}, \mathcal{D}_n). \tag{1}$$

where $\mathcal{D} = \cup_{n=1}^{N} \mathcal{D}_n$ can be regarded as the dataset aggregated from all the mobile users and $\ell(\boldsymbol{w}; \boldsymbol{x}_i, y_i)$ is the empirical loss associated with the i-th data sample pair $(\boldsymbol{x}_i, y_i)$ and $\alpha_n \in [0, 1]$ denotes the weight assigned on the data set of user n, satisfying $\sum_{n=1}^{N} \alpha_n = 1$; here, $F_n(\boldsymbol{w}, \mathcal{D}_n)$ represents the local loss function constructed by the dataset of the mobile user n, given as

$$F_n(\boldsymbol{w}, \mathcal{D}_n) = \frac{1}{|\mathcal{D}_n|} \sum_{(\boldsymbol{x}_i, y_i) \in \mathcal{D}_n} \ell(\boldsymbol{w}; \boldsymbol{x}_i, y_i). \tag{2}$$

Toward minimizing (1), the network employs FL to train the models. The training is constituted by a sequence of edge intervals (within which edge servers interact with mobile users) and cloud intervals (where the cloud server interacts with the edge servers). The edge and cloud intervals comprise several consecutive communication rounds and edge intervals, respectively.

B. Federated Model Training

This part details the concrete steps in training the model.

1) Model Initialization of Cloud Server: At the beginning of a cloud interval, the cloud server distributes the initial cloud model parameter to each edge server.

2) Model Initialization of Edge Server: At the beginning of an edge interval, each edge server broadcasts the initial edge model to every connected user. The set of mobile users served by the m-th edge server during the p-th edge interval is denoted by $\mathcal{N}_m^p$. Note that as the users roam across the network, the set $\mathcal{N}_m^p$, $m \in \{1, ..., M\}$, varies over time.

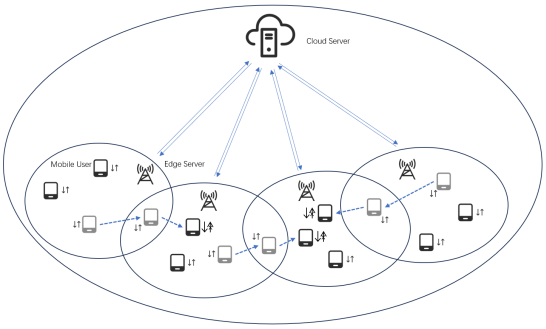

Fig. 1: Network architecture.

3) Local Updates: After receiving the initial model from the edge server, each mobile user performs τ_{e} rounds of SGD iteration using its local dataset. Such an operation at the typical user n can be formally expressed as follows:

$$\boldsymbol{w}_n^{t+1} \leftarrow \boldsymbol{w}_n^t - \eta g_n(\boldsymbol{w}_n^t, \xi_n^t). \tag{3}$$

where η is the learning rate, $\xi_n^t \subseteq \mathcal{D}_n$ represents the mini-batch IID sampled from local dataset during the t-th local updates, and $g_n(w_n^t, \xi_n^t)$ denotes the (stochastic) gradient of local loss function evaluated based on the mini-batch dataset. In this network, each user keeps moving while updating its local model, leading to a dynamic connectivity status with the edge servers.

4) Edge Aggregation: Upon finishing the local training, each user tries to upload the resultant model parameters back to the edge server from which it received the initial parameter. For a generic edge server m, we consider only users still connected with it can successfully send their local model back, which we refer to as the *participating users*. Note that the time of the p-th edge aggregation is also the time of model initialization for the $(p+1)$-th edge interval. As such, the participating users of the p-th edge interval of the m-th edge server can be formally written as $\mathcal{N}_m^p \cap \mathcal{N}_m^{p+1}$, which we denote by $\mathcal{S}_m^p$. Then, the edge server aggregates the model parameters from mobile users (still) in its coverage, as follows:

$$\bar{\boldsymbol{w}}_m^{p\tau_{\mathrm{e}}} \leftarrow \frac{1}{|\mathcal{S}_m^p|} \sum_{n \in \mathcal{S}_m^p} \boldsymbol{w}_n^{p\tau_{\mathrm{e}}}. \tag{4}$$

After updating the model parameters, each edge server distributes the aggregated result to its connected users for the next τ_{e} round local updates.

5) Cloud Aggregation: After τ_{c} rounds of edge aggregations, the cloud server collects the aggregation results from each edge server and averages it as follows:

$$\bar{\boldsymbol{w}}^{c\tau_{\mathrm{c}}\tau_{\mathrm{e}}} \leftarrow \sum_{m=1}^{M} \frac{|\mathcal{N}_m^{c\tau_{\mathrm{c}}}|}{|\mathcal{N}^{c\tau_{\mathrm{c}}}|} \bar{\boldsymbol{w}}_m^{c\tau_{\mathrm{c}}\tau_{\mathrm{e}}}. \tag{5}$$

After the cloud aggregation, the cloud server then sends the result back to the edge server for model initialization.

C. User Mobility Model

User mobility primarily impacts the dynamics of connection status with the edge servers. We adopt the Markov chain model to characterize this feature. Under this model, every user will stay or move to a neighboring edge server during a time slot with a certain probability.

Specifically, let $\boldsymbol{s}_n^t \in \mathbb{R}^M$ be a vector denoting the connection status of the n-th mobile users at the beginning of the t-th time slot, where the entries are defined as follows:

$$\boldsymbol{s}_n^t[m] = \begin{cases} 1, & \text{if } n \in \mathcal{N}_m^t, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

in which $\sum_{m=1}^{M} \boldsymbol{s}_n^t[m] = 1, \forall n, t$ since each user is only connected with only one edge server at a time.

In this paper, we assume all the mobile users adhere to a common transition probability matrix $P \in [0,1]^{M\times M}$, where $p_{ij} \in [0,1]$ denotes the probability that a typical user moves from the i-th edge server to j-th edge server (particularly, p_{ii} denotes the probability of staying on i-th edge server). We require $\sum_{j=1}^{M} p_{ij} = 1$, since each user either roams to another server or stays in the original cell during a communication round. Given the current observation vector and transmission probability matrix, the future state vector of n-th user can be calculated by

$$\boldsymbol{s}_n^{t+1} = \boldsymbol{s}_n^t P^t = \boldsymbol{s}_n^0 \prod_{\tau=0}^{t} P^\tau. \tag{7}$$

Assuming the transition matrix P is irreducible, then there exists a steady-state distribution $\boldsymbol{s}$ satisfying $\boldsymbol{s} = \boldsymbol{s}P$. Following the above, we can calculate the size N_m^t of the set $\mathcal{N}_m^t$ as follows

$$N_m^t = \sum_{n=1}^{N} \boldsymbol{s}_n^t[m]. \tag{8}$$

which gives

$$\sum_{n=1}^{N} \boldsymbol{s}_n[m] = \boldsymbol{s}[m]N, \quad \text{as } t \to \infty. \tag{9}$$

We assume the network has reached the steady state, where all the mobile users are uniformly and randomly distributed over the entire network following the steady-state distribution, $\boldsymbol{s}$, at the beginning of each time, and the average number of incoming and departing users of each edge server will be balanced, i.e., $N_m^t = N_m = \boldsymbol{s}[m]N, \forall m, \forall t$. As such, we can assume the number of users of each edge server follows its expectation. Note that the indices of users in $\mathcal{N}_m^t$ are time-varying, albeit the size remains the same in expectation.

D. Illustrate the impact of Mobile Users

In this part, we illustrate influence of user mobility and why a higher user mobility will bring about a drastic drop to the algorithm performance in a Non-IID scenario.

To make the illustration more clear, we introduce the auxiliary virtual sequence as [12] to represent the intermediate result after one-step SGD from $\bar{\boldsymbol{w}}^t$ and interpret $\bar{\boldsymbol{w}}^{t+1}$ as the parameter obtained after communication steps (if there exists). Then the model parameter of n-th user connected with m-th edge server within p-th edge interval can be described as follows:

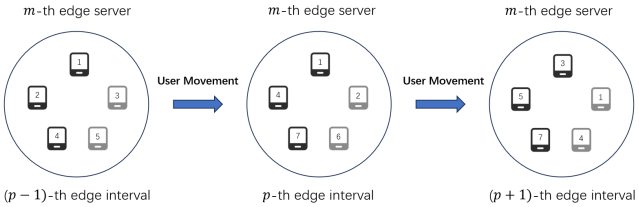

Fig. 2: The participating user of m-th edge server

$$\boldsymbol{v}_n^{t+1} = \boldsymbol{w}_n^t - \eta \nabla F_n(\boldsymbol{w}_n^t, \xi_n^t), \tag{10}$$

$$\boldsymbol{w}_n^{t+1} = \begin{cases} \boldsymbol{v}_n^{t+1}, & \text{if } \tau_{\mathrm{e}} \nmid t+1, \\ \frac{1}{|\mathcal{S}_m^p|} \sum_{n\in\mathcal{S}_m^p} \boldsymbol{v}_n^{t+1}, & \text{if } \tau_{\mathrm{e}} \mid t+1, \tau_{\mathrm{c}}\tau_{\mathrm{e}} \nmid t+1, \\ \frac{1}{|\mathcal{S}^p|} \sum_{n\in\mathcal{S}^p} \boldsymbol{v}_n^{t+1}, & \text{if } \tau_{\mathrm{c}}\tau_{\mathrm{e}} \mid t+1. \end{cases} \tag{11}$$

in which $\mathcal{S}^p = \cup_{m=1}^{M} \mathcal{S}_m^p$ denotes the participating users for p-th edge aggregation.

And we introduce two virtual sequences $\bar{\boldsymbol{w}}_m^t = \frac{1}{|\mathcal{N}_m^p|} \sum_{n\in\mathcal{N}_m^p} \boldsymbol{w}_n^t$ and $\bar{\boldsymbol{v}}_m^t = \frac{1}{|\mathcal{N}_m^p|} \sum_{n\in\mathcal{N}_m^p} \boldsymbol{v}_n^t$ for each m-th edge server within p-th edge interval. Then $\bar{\boldsymbol{v}}_m^{t+1}$ results from a single step SGD from $\bar{\boldsymbol{w}}_m^t$ and then $\bar{\boldsymbol{w}}_m^{t+1}$ is obtained from $\bar{\boldsymbol{v}}_m^{t+1}$ after potential aggregation.

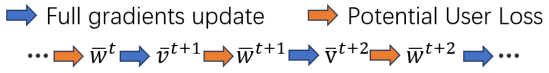

Fig. 3: Illustration of the update of $\bar{\boldsymbol{w}}_m^t$ and $\bar{\boldsymbol{v}}_m^t$ for m-th edge server

When mobility is considered, user may move in and out of the coverage area of edge server randomly. Thus the set of users connected with m-th edge server at the beginning and end of the edge interval differs as showed in Fig. 2, which leads to a decrease for the number of participating users compared with situation without mobile users. As shown in Fig. 3, $\bar{\boldsymbol{w}}_m^t$ and $\bar{\boldsymbol{v}}_m^t$ keep the same in the interior of each edge interval, while at the end of the p-th edge interval, $\bar{\boldsymbol{w}}_m^t$ and $\bar{\boldsymbol{v}}_m^t$ differ due to the user loss of the edge aggregation. Later, we will show that this difference is caused by two parts: connection loss from user mobility and SGD noises, which interact in a multiplicative manner. And with some assumptions, we can bound this difference.

III. Convergence Analysis

In this section, we analyze the convergence rate of the considered edge training system. To facilitate the analysis, we first make the following assumptions, which are commonly adopted in literature [15].

Assumption 1 (Lipschitz gradient): *There exists a constant L such that*

$$\|\nabla F_n(\boldsymbol{w}_1) - \nabla F_n(\boldsymbol{w}_2)\| \leq L\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|, \forall n, \boldsymbol{w}_1, \boldsymbol{w}_2. \quad (12)$$

Assumption 2 (Bounded Stochastic Gradient Noise): *The stochastic gradient of all users has a uniform upper bound:*

$$\mathbb{E}_{\xi_n^t \sim \mathcal{D}_n}\left[\|g_n(\boldsymbol{w}; \xi_n^t) - \nabla F_n(\boldsymbol{w})\|^2\right] \leq \sigma^2, \forall n, \boldsymbol{w}. \quad (13)$$

Assumption 3 (Bounded Upward and Downward Divergence): *The gradient divergence of the m-th edge server and the cloud server can be respectively bounded as follows:*

$$\sum_{n \in \mathcal{N}_m^p} \frac{1}{N_m} \|\nabla F_n(\boldsymbol{w}) - \nabla f_m(\boldsymbol{w})\|^2 \leq \epsilon_m^2, \quad \forall i, \boldsymbol{w}. \quad (14)$$

$$\sum_{m=1}^{M} \frac{N_m}{N} \|\nabla f_m(\boldsymbol{w}) - \nabla f(\boldsymbol{w})\|^2 \leq \epsilon^2, \quad \forall \boldsymbol{w}. \quad (15)$$

where f_m denotes the loss function of the m-th edge server.

In addition, for the m-th edge server and p-th edge interval, we denote by $\mathcal{S}_m^p$ and $\mathcal{N}_m^p$ the set of staying users and users at the beginning of p-th edge interval, respectively; and let $K_m = |\mathcal{S}_m^p|, N_m = |\mathcal{N}_m^p|$ be the size of two sets.

Based on the above assumptions, we commence our analysis by bounding the difference in the aggregated global parameters, taking into account the effects of user mobility.

Lemma 1: *We have*

$$\mathbb{E}_{\mathcal{S}_m^p}[\bar{\boldsymbol{w}}_m^{p\tau_\mathrm{e}}] = \bar{\boldsymbol{v}}_m^{p\tau_\mathrm{e}}. \quad (16)$$

And we can bound the variance as:

$$\mathbb{E}_{\mathcal{S}_m^p}\left[\|\bar{\boldsymbol{w}}_m^{p\tau_\mathrm{e}} - \bar{\boldsymbol{v}}_m^{p\tau_\mathrm{e}}\|^2\right] = \frac{1}{N_m(N_m-1)}\left(\frac{1}{p_m} - 1\right) \sum_{n \in \mathcal{N}_m^p} \|\boldsymbol{v}_n^{p\tau_\mathrm{e}} - \bar{\boldsymbol{v}}_m^{p\tau_\mathrm{e}}\|^2. \quad (17)$$

in which $p_m = K_m/N_m$ denotes the staying probability for each user in m-th edge server.

Proof: See Appendix A in the supplementary material. ■

The above lemma indicates that the divergence induced by user mobility is affected by the staying probability p_m and the mean square error (MSE), $\sum_{n \in \mathcal{N}_m^p} \|\boldsymbol{v}_n^{p\tau_\mathrm{e}} - \bar{\boldsymbol{v}}_m^{p\tau_\mathrm{e}}\|^2$, of user parameters of one edge server within one edge interval. Moreover, user mobility acts as an amplifying factor that enlarges the divergence in local updates: the higher the user mobility, the larger this divergence.

Next, we bound the expected gradient norm as [8], [15].

Lemma 2: *Let $\eta \leq \frac{1}{L}$ and $T = C\tau_\mathrm{c}\tau_\mathrm{e}$, then we have*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\bar{\boldsymbol{w}}^t)\|^2\right] \leq{}& \frac{2\left(\mathbb{E}\left[f(\bar{\boldsymbol{w}}^0)\right] - f^*\right)}{\eta T} + \frac{\eta L \sigma^2}{N} \\ &+ 2L^2\Bigg(\frac{1}{T}\sum_{t=0}^{T-1}\sum_{m=1}^{M}\frac{N_m}{N}\cdot\frac{1}{N_m}\sum_{n\in\mathcal{N}_m^p}\mathbb{E}\left[\|\bar{\boldsymbol{w}}_m^t - \boldsymbol{w}_n^t\|^2\right] \\ &\quad + \frac{1}{T}\sum_{t=0}^{T-1}\sum_{m=1}^{M}\frac{N_m}{N}\mathbb{E}\left[\|\bar{\boldsymbol{w}}^t - \bar{\boldsymbol{w}}_m^t\|^2\right]\Bigg) \\ &+ \frac{L}{\eta T}\sum_{p=1}^{C\tau_\mathrm{c}}\mathbb{E}\left[\|\bar{\boldsymbol{w}}^{p\tau_\mathrm{e}} - \bar{\boldsymbol{w}}^{p\tau_\mathrm{e}}\|^2\right].\end{aligned} \quad (18)$$

Proof: See Appendix B in the supplementary material. ■

Compared with the result in [15], the Lemma 2 has an extra term (namely, the last term on the right-hand side of the above inequality), which is attributed to the connectivity loss incurred by user mobility.

According to Lemma 2, the convergence rate is determined by the MSE of the edge server's aggregated parameters, edge parameters, and user mobility. We bound these quantities respectively in the sequel.

First, we bound the MSE of the user parameters aggregated at each edge server.

Lemma 3: *The MSE of the m-th edge server's user parameters can be bounded as:*

$$\frac{1}{T}\sum_{t=1}^{T-1}\frac{1}{N_m}\sum_{n\in\mathcal{N}_m^p}\mathbb{E}\left[\|\bar{\boldsymbol{w}}_m^t - \boldsymbol{w}_n^t\|^2\right] \leq \frac{\tau_\mathrm{e} A_m}{1 - 12\eta^2 L^2 \tau_\mathrm{e}^2}. \quad (19)$$

in which

$$A_m = 2\eta^2\left(1 - \frac{1}{N_m}\right)\sigma^2 + 6\eta^2\epsilon_m^2\tau_\mathrm{e}. \quad (20)$$

Proof: See Appendix C in the supplementary material. ■

Notably, downward MSE is not influenced by user mobility, since in our model, user movement is only observable when the edge aggregation happens, that is, at the end of each edge interval. Thus within an edge interval, there's no impact of user mobility.

Lemma 4: *The MSE of edge parameters can be bounded as:*

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\sum_{m=1}^{M}\frac{N_m}{N}\mathbb{E}\left[\|\bar{\boldsymbol{w}}^t - \bar{\boldsymbol{w}}_m^t\|^2\right] \\ &\leq \frac{\eta^2 B}{1 - 12\eta^2\tau_\mathrm{c}^2\tau_\mathrm{e}^2 L^2} + \frac{4\eta^2\tau_\mathrm{c}^2\tau_\mathrm{e}^2 L^2}{1 - 12\eta^2\tau_\mathrm{c}^2\tau_\mathrm{e}^2 L^2}\cdot\frac{\tau_\mathrm{e} A}{1 - 12\eta^2\tau_\mathrm{e}^2 L^2}.\end{aligned} \quad (21)$$

in which

$$A = \sum_{m=1}^{M}\frac{N_m}{N}A_m = 2\eta^2\sigma^2\left(1 - \frac{M}{N}\right) + 6\eta^2\tau_\mathrm{e}\sum_{m=1}^{M}\frac{N_m}{N}\epsilon_m^2, \quad (22)$$

$$B = 4\tau_\mathrm{c}\tau_\mathrm{e}\sigma^2\frac{1}{N}\sum_{m=1}^{M}\frac{1}{p_m}\left(1 - \frac{N_m}{N}\right) + 6\tau_\mathrm{c}^2\tau_\mathrm{e}^2\epsilon^2. \quad (23)$$

Proof: See Appendix D in the supplementary material. ■

Here, the MSE of edge parameters is influenced by user mobility, but only the term with stochastic gradient noise is enlarged.

Lemma 5: *Let $T = C\tau_\mathrm{c}\tau_\mathrm{e}$, the divergence of user mobility can be bounded as:*

$$\begin{aligned}&\frac{1}{T}\sum_{p=1}^{C\tau_\mathrm{c}}\mathbb{E}\left[\|\bar{\boldsymbol{w}}^{p\tau_\mathrm{e}} - \bar{\boldsymbol{w}}^{p\tau_\mathrm{e}}\|^2\right] \\ &\leq \frac{1}{1 - 12\eta^2 L^2\tau_\mathrm{e}^2}\sum_{m=1}^{M}\frac{N_m}{N}\left(\frac{1}{p_m} - 1\right)\frac{A_m}{N_m - 1}.\end{aligned} \quad (24)$$

Proof: See Appendix E in the supplementary material. ■

Remark 1: *This lemma discloses that the divergence of user mobility stems from two aspects, i.e., the stochastic gradient noise and data heterogeneity. Moreover, these effects are enlarged by a factor related to the staying probability. As data heterogeneity increases, downward divergence goes up, as well as this divergence. Thus, this divergence increases with both user mobility and data heterogeneity. This may explain why a significant decline of performance occurs when HFL meets high user mobility and high data heterogeneity. Specifically, when $p_m = 1$, this term goes to zero, which is the situation without user mobility.*

Piecing together all the above lemmas, we can obtain the final convergence result of HFL with mobile users.

Theorem 1: *Under the employed HFL system, by choosing learning rate as η satisfying $\eta \leq \frac{1}{2\sqrt{6}L\tau_c\tau_e}$, we have*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla f(\bar{\boldsymbol{w}}^t)\|^2\big] &\leq \frac{2\big(\mathbb{E}\big[f(\bar{\boldsymbol{w}}^0)\big]-f^*\big)}{\eta T}+\frac{\eta L\sigma^2}{N}\\ &+\frac{4\eta L}{N}\sum_{m=1}^{M}\Big(\frac{1}{p_m}-1\Big)\Big(\sigma^2+3\tau_e\frac{N_m}{N_m-1}\epsilon_m^2\Big)\\ &+8\eta^2L^2\Big(\frac{2\tau_c\tau_e\sigma^2}{N}\sum_{m=1}^{M}\frac{1}{p_m}\Big(1-\frac{N_m}{N}\Big)+3\tau_c^2\tau_e^2\epsilon^2\\ &+\frac{4}{3}\sigma^2\Big(1-\frac{M}{N}\Big)\tau_e+4\sum_{m=1}^{M}\frac{N_m}{N}\epsilon_m^2\Big).\end{aligned} \tag{25}$$

Proof: See Appendix F in the supplementary material. ■

Theorem 1 encompasses several key system factors, including user mobility, data heterogeneity, and hierarchical network topology. It shows that by adequately adjusting the step size, after a sufficient amount of model training, the model converges toward a region around the optimality.

IV. Experiments

In this section, we carry out experiments to validate our theoretical results. Specifically, we consider training a convolutional neural network (CNN) over the CIFAR-10 data set [17] with non-IID data partitioning across users. We consider the HFL system containing 50 users and 5 edge servers. All users are randomly assigned to the edge servers based on a stable distribution $\boldsymbol{s}$ and subsequently move according to our defined mobility model. The mini-batch size at each SGD step is set as 100, and the learning rate is set as 0.1. We perform 50 cloud communication rounds in total and set edge interval length $\tau_e = 10$, cloud interval length $\tau_c = 2$.

To characterize data heterogeneity, we explore the non-IID setting introduced in [1]. In particular, a pathological non-IID scenario where distinct shards for each user represent varying degrees of non-IID. The experiments are conducted under 2 shards, 3 shards, 5 shards, and IID, respectively.

Fig. 4 demonstrates the influence of the user staying probability, p, on the convergence rate of the system. Each curve represents the average accuracy obtained from multiple experiments, with the shaded area indicating the accuracy range across the experiments. This figure confirms that user mobility significantly affects the convergence rate in all scenarios. Particularly, as data heterogeneity increases, the performance degradation becomes more pronounced. Additionally, we observe more fluctuations in the test accuracy with an increase in user mobility and/or data heterogeneity.

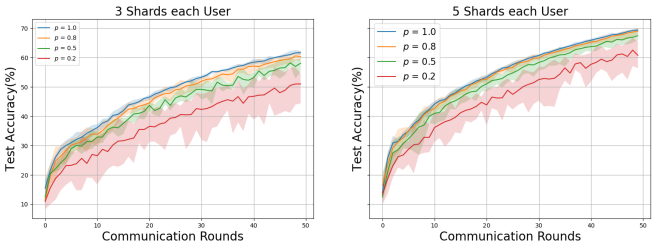

Fig. 4: Test accuracy under different data heterogeneity with various staying probability.

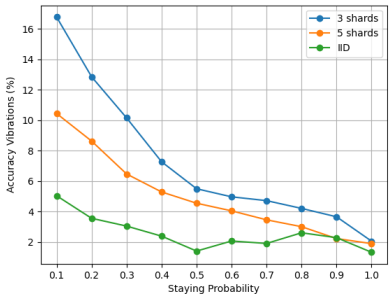

Fig. 5: Accuracy vibration versus staying probability.

Fig. 5 plots the vibration range of final accuracy as a function of staying probability under different degrees of data heterogeneity. We can see that for each non-IID scenario, the vibration range increases with decreasing staying probability, which is in line with the observations from our analysis. Furthermore, the decrease in staying probability accentuates the difference in vibration range across various data heterogeneity. This indicates that a small staying probability magnifies the divergence arising from data heterogeneity, coinciding with the analysis in Remark 1.

V. Conclusion

We investigated the interplay between user mobility and data heterogeneity in an HFL system. We leveraged a Markov chain to model user mobility, capturing the impact of user mobility on the updates of intermediate parameters. Our analysis quantifies the interplay between user mobility and data heterogeneity, showing that the degradation intensifies with increased levels of user mobility and data heterogeneity. Consequently, we derived a convergence rate of the HFL, encompassing effects of user mobility, data heterogeneity, and system hierarchy. The analysis does not assume convexity of the objective function and hence applies to even the deep learning settings. Our theoretical findings offer valuable insights into the dynamics of HFL systems with mobile users, and experiments validate our results.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Stat. (AISTATS)*, vol. 54, Fort Lauderdale, USA, Apr. 2017, pp. 1273–1282.

[2] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 2031–2063, 3rd Quart. 2020.

[3] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[4] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proc. IEEE*, vol. 107, no. 11, pp. 2204–2239, Nov 2019.

[5] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Client-edge-cloud hierarchical federated learning," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Dublin, Ireland, June 2020, pp. 1–6.

[6] T. Castiglia, A. Das, and S. Patterson, "Multi-level local sgd: Distributed sgd for heterogeneous hierarchical networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2020.

[7] S. Luo, X. Chen, Q. Wu, Z. Zhou, and S. Yu, "HFEL: Joint edge association and resource allocation for cost-efficient hierarchical federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6535–6548, Oct. 2020.

[8] C. Feng, H. H. Yang, D. Hu, Z. Zhao, T. Q. S. Quek, and G. Min, "Mobility-aware cluster federated learning in hierarchical wireless networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 8441–8458, Apr. 2022.

[9] J. Park, S. Samarakoon, A. Elgabli, J. Kim, M. Bennis, S.-L. Kim, and M. Debbah, "Communication-efficient and distributed learning over wireless networks: Principles and applications," *Proc. IEEE*, vol. 109, no. 5, pp. 796–819, May 2021.

[10] J. Kang, Z. Xiong, D. Niyato, Y. Zou, Y. Zhang, and M. Guizani, "Reliable federated learning for mobile networks," *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 72–80, Apr. 2020.

[11] Y. Peng, X. Tang, Y. Zhou, Y. Hou, J. Li, Y. Qi, L. Liu, and H. Lin, "How to tame mobility in federated learning over mobile networks?" *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 9640–9657, Dec. 2023.

[12] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Dec 2019.

[13] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *Available at arXiv:1806.00582*, 2018.

[14] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated learning on non-iid data silos: An experimental study," in *Proc. Int. Conf. on Data Eng. (ICDE)*, Kuala Lumpur, Malaysia, May 2022.

[15] J. Wang, S. Wang, R.-R. Chen, and M. Ji, "Demystifying why local aggregation helps: Convergence analysis of hierarchical sgd," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 36, no. 8, Jun. 2022, pp. 8548–8556.

[16] M. Chen, H. V. Poor, W. Saad, and S. Cui, "Performance optimization of federated learning over mobile wireless networks," in *Proc. IEEE Int. Workshop Signal Process. Adcances Wireless Commun.*, Virtual, May 2020, pp. 1–5.

[17] A. Krizhevsky, "Learning multiple layers of features from tiny images," *Tech. Rep. (TR-2009)*, 2009.