Multi-Stage Hybrid Federated Learning over Large-Scale D2D-Enabled Fog Networks

Seyyedali Hosseinalipour, *Member, IEEE*, Sheikh Shams Azam, Christopher G. Brinton, *Senior Member, IEEE*, Nicolò Michelusi, *Senior Member, IEEE*, Vaneet Aggarwal, *Senior Member, IEEE*, David J. Love, *Fellow, IEEE*, and Huaiyu Dai, *Fellow, IEEE*

***Abstract*—Federated learning has generated significant interest, with nearly all works focused on a "star" topology where nodes/devices are each connected to a central server. We migrate away from this architecture and extend it through the *network* dimension to the case where there are multiple layers of nodes between the end devices and the server. Specifically, we develop multi-stage hybrid federated learning (MH-FL), a hybrid of intra- and inter-layer model learning that considers the network as a *multi-layer cluster-based structure*. MH-FL considers the *topology structures* among the nodes in the clusters, including local networks formed via device-to-device (D2D) communications, and presumes a *semi-decentralized architecture* for federated learning. It orchestrates the devices at different network layers in a collaborative/cooperative manner (i.e., using D2D interactions) to form *local consensus* on the model parameters and combines it with multi-stage parameter relaying between layers of the tree-shaped hierarchy. We derive the upper bound of convergence for MH-FL with respect to parameters of the network topology (e.g., the spectral radius) and the learning algorithm (e.g., the number of D2D rounds in different clusters). We obtain a set of policies for the D2D rounds at different clusters to guarantee either a finite optimality gap or convergence to the global optimum. We then develop a distributed control algorithm for MH-FL to tune the D2D rounds in each cluster over time to meet specific convergence criteria. Our experiments on real-world datasets verify our analytical results and demonstrate the advantages of MH-FL in terms of resource utilization metrics.**

***Index Terms*—Fog learning, device-to-device communications, peer-to-peer learning, cooperative learning, distributed machine learning, semi-decentralized federated learning.**

I. Introduction

Machine learning (ML) has produced automated solutions to problems ranging from natural language processing to object detection/tracking [1], [2]. Traditionally, ML model training has been carried out at a central node (e.g., a server). In many contemporary applications of ML, however, the relevant data is generated at the end user devices. As these devices generate larger volumes of data, transferring it to a central server for model training has several drawbacks: (i) it may require significant energy consumption from battery-powered devices; (ii) the round-trip-times between data generation and model training may incur prohibitive delays; and (iii) in privacy-sensitive applications, end users may not be willing to transmit their raw data in the first place.

S. Hosseinalipour, S. Azam, C. Brinton, V. Aggarwal, and D. Love are with Purdue University: {hosseina,azam1,cgb,vaneet,djlove}@purdue.edu. N. Michelusi is with Arizona State University: nicolo.michelusi@asu.edu. H. Dai is with NC State University: hdai@ncsu.edu.

C. Brinton was supported in part by ONR under grant N00014-21-1-2472, and NSC grant W15QKN-15-9-1004. Part of Michelusi's research has been funded by NSF under grants CNS-1642982 and CNS-2129015. D. Love was supported in part by the NSF under grant EEC1941529. H. Dai was supported by NSF CNS-1824518.

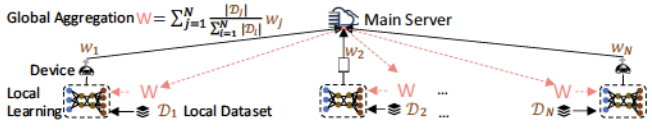


Fig. 1: Conventional *star* topology architecture of federated learning.

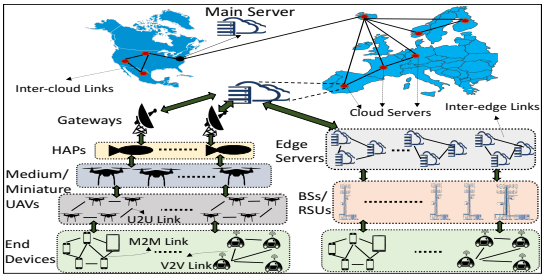
Federated learning has emerged as a technique for distributing model training across devices while keeping the devices' datasets local [3]. Its conventional architecture consists of a main server connected to multiple devices in a *star* topology (see Fig. 1). Each round of model training consists of two steps: (i) *local updating*, where each device updates its local model based on its dataset and the global model, e.g., using gradient descent, and (ii) *global aggregation*, where the server gathers devices' local models and computes a new global model, which is then synchronized across the devices to begin the next round.

In conventional federated learning, only device-to-server (in step (i)) and server-to-device (in step (ii)) communications occur. This is limiting and prohibitive in contemporary large-scale network scenarios, where there are several layers of nodes between the end devices and the cloud (see Fig. 2(a)). In particular, it can lead to long delays, large bandwidth utilization, and high power consumption for aggregations [4]. We migrate federated learning from its star structure to a more distributed structure that accounts for the multi-layer network dimension and leverages *topology structures* among the devices.

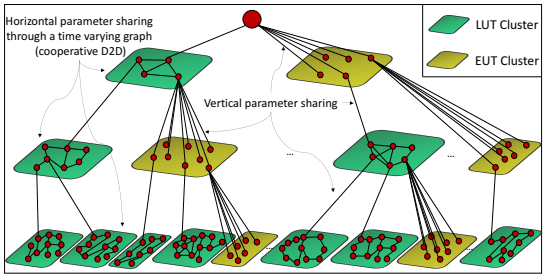
A. Fog Learning: Federated Learning in Fog Environments

Fog computing is an emerging technology which aims to manage computation resources across the cloud-to-things continuum, encompassing the cloud, core, edge, metro, clients, and things [5]. We recently introduced the fog learning (FogL) paradigm [4], which advocates leveraging the fog computing architecture to handle ML tasks. Specifically, FogL requires extending federated learning to (i) incorporate fog network structures, (ii) account for device computation heterogeneity, and (iii) manage the proximity between resource-abundant and resource-constrained nodes. Our focus in this paper is (i), i.e., extending federated learning along the *network* dimension.

We consider the sample network architecture of FogL in Fig. 2(a). There are multiple layers between user devices and the cloud, including base stations (BSs) and edge servers. Compared with federated learning, FogL has two key characteristics: (i) It assumes a multi-layer cluster-based structure with local parameter aggregations at different layers. (ii) In addition to inter-layer communications, it includes intra-layer communications via device-to-device (D2D) connectivity,



(a): A schematic of model transfer stages for a large-scale ML task in a fog network. The parameters of the end devices are carried through multiple layers of the network consisting of base stations (BSs), road side units (RSUs), unmanned aerial vehicles (UAVs), high altitude platforms (HAPs), edge servers, and cloud servers before reaching the main server. Devices located at different layers of the network can engage in direct communications via mobile-mobile (M2M), vehicle-vehicle (V2V), UAV-UAV (U2U), inter-edge, and inter-cloud links.



(b): Partitioning the network layers into multiple LUT/EUT clusters for FogL, introducing a hybrid model training framework consisting of both horizontal and vertical parameter transfer. Inside each LUT cluster, the devices engage in collaborative/cooperative D2D communications through a time varying network topology and exchange their parameters. The parents of LUT clusters obtain the consensus of their children parameters by sampling the model parameter of one node, while the parents of EUT clusters receive all the parameters of their children.

Fig. 2: Architecture of (a) the multi-layer network structure of fog computing systems and (b) the network layers and parameters transfer for FogL.

which is promoted in 5G and IoT [6]. There exists a literature on D2D communication protocols for ad-hoc and sensor networks including vehicular ad-hoc networks (VANET), mobile ad-hoc networks (MANET), and flying ad-hoc networks (FANET) [7]–[10]. Exploiting D2D communications has also been promoted in agriculture and rural use cases [11], making FogL a promising model training strategy in such environments. FogL also considers server-to-server interactions [12] and other types of peer-to-peer (P2P) interactions under the umbrella of D2D.

Characteristic (ii) mentioned above orchestrates the devices at each layer in a cooperative framework, introducing a set of *local networks* to the learning paradigm. This motivates studying the learning performance with explicit consideration of *topology* structure among the devices. To do this, we partition network layers into clusters of two types as depicted in Fig. 2(b): (i) limited uplink transmission (LUT), where D2D communications are enabled, and (ii) extensive uplink transmission (EUT), where, similar to the conventional federated learning, all nodes only communicate with their upper layer.

Accommodating both inter- and intra-layer communications introduces a *semi-decentralized learning architecture*, which is a *hybrid* model for training that considers conventional server-device interactions (i.e., centralized "star" topology) in conjunction with collaborative/cooperative D2D communications (i.e., fully decentralized "mesh" topology). Thus, the methodology we develop in this paper is called *multi-stage hybrid federated learning* (`MH-FL`) and considers both intra-cluster consensus formation and inter-cluster aggregations for distributed ML. In developing `MH-FL`, we incorporate the time-varying local network topologies among the devices as a dimension of federated learning, and demonstrate how it impacts ML model convergence and accuracy.

B. Related Work

Researchers have considered the effects of limited and imperfect communication capabilities in wireless networks – such as channel fading, packet loss, and limited bandwidth – on the operation of federated learning [13]–[15]. Also, communication techniques such as quantization [16], [17] and sparsification of model updates (i.e., when only a fraction of the model parameters are shared during model training) [18] have been studied. Recently, [19] analyzed the convergence bounds in the presence of edge network resource constraints.

Research has also considered the computation aspects of federated learning in wireless networks [14], [20]–[22]. Part of this literature has focused on learning in the presence of *stragglers*, i.e., when a node has significantly lower computation capabilities than others [14], [20]. Another emphasis has been reducing the computation requirements through intelligent raw data offloading between devices [21] and judicious selection of device participation [22], [23]. Other techniques for mitigating compute limitations, e.g., through model compression, have also been applied to distributed ML [24].

There exist recent works on hierarchical federated learning [25]–[27]. These works are mainly focused on specific use cases of two-tiered network structures above wireless cellular devices, e.g., edge clouds connected to a main server [25], [26] or small cell and macro cell base stations [27]. As compared to all the aforementioned works, which consider the star model training topology (or *tree* in case of hierarchical considerations), our work is distinct for several reasons, including that: (i) we introduce a *multi-layer cluster-based structure* with an arbitrary height that encompasses all IoT elements between the end devices and the main server, which generalizes all the prior models; and (ii) more importantly, we explicitly consider the *network* dimension and topology structure among the devices at each network layer formed via cooperative/collaborative D2D communications. This migrates us from prior models and enables new analysis and considerations for hybrid intra- and inter-layer model learning over large-scale fog networks.

There also exist recent works on *fully decentralized (serverless)* federated learning [17], [28], [29]. These architectures require a well-connected D2D communication graph among all the devices in the network, which becomes less feasible to maintain as the geographical span of the devices increases (e.g., the end devices across multiple regions in Fig. 2(a)). We establish an intermediate learning architecture that couples the star topology assumed in conventional federated learning with fully decentralized architectures to provide a *scalable model training*. In particular, we propose a novel semi-decentralized learning architecture that (i) uses a cluster-based representation of devices with local communications only among the D2D-enabled devices inside the same cluster; (ii) reduces the reliance on resource-intensive uplink model transmissions via sampling only one device from D2D-enabled clusters; and (iii) is based

on a layered coordination of global aggregations across the fog learning hierarchy facilitated by a main server.

Beyond federated learning, there is a well developed literature on other distributed ML techniques (e.g., [30]–[32]). Our proposed framework for FogL inherits its model aggregation rule from federated learning, i.e., local gradient descent at the devices and weighted averaging to obtain the global model. We choose this due to specific characteristics that make it better suited for fog: keeping the user data local, handling non-iid datasets across devices, and handling imbalances between sizes of local datasets [3]. These capabilities have made federated learning the most widely acknowledged distributed learning framework for future wireless systems [33], [34]. The multi-stage hybrid architecture of FogL could also be studied in the context of other distributed ML techniques, e.g., ADMM [30].

Finally, there exist a literature on distributed consensus with applications in multi-agent systems [35], [36], sensor networks [37], [38], and optimization [39]–[42]. Our scenario is unique given its multi-layer network structure and focus on a hybrid ML model training, where the goal is to propagate an expectation of the nodes' parameters through the hierarchy to train an ML model. The results we obtain have thus not yet appeared in either consensus-related or ML-related literature.

C. Summary of Contributions

Our contributions in this work can be summarized as follows:

- We formalize multi-stage hybrid federated learning (MH−FL), a new methodology for distributed ML. MH−FL extends federated learning along the network dimension, relaying the local updates of end devices through the network hierarchy via a novel multi-stage, cluster-based parameter aggregation technique. This paradigm introduces local aggregations achieved by an interplay between cooperative D2D communications and distributed consensus formation.
- We analytically characterize an upper bound of convergence of MH−FL. We demonstrate how this bound depends on characteristics of the ML model, the network topology, and the learning algorithm, including the number of model parameters, the communication graph structure, and the number of D2D rounds at different device clusters.
- We demonstrate that the model loss achieved by MH−FL under unlimited D2D rounds coincides with that of federated learning. Under the finite D2D rounds regime, we obtain a condition under which a constant optimality gap can be achieved asymptotically. We further show that under limited finely-tuned D2D rounds, MH−FL converges linearly to the optimal ML model. We further introduce a practical cluster sampling technique and investigate its convergence behavior.
- We obtain analytical relationships for tuning (i) the number of D2D rounds in different clusters at different layers and (ii) the number of global iterations to meet certain convergence criteria. We use these relationships to develop distributed control algorithms that each cluster can employ individually to adapt the number of D2D rounds it uses over time.
- Our experimental results on real-world datasets verify our theoretical findings and show that MH−MT can improve network resource utilization significantly with negligible impact on model training convergence speed and accuracy.

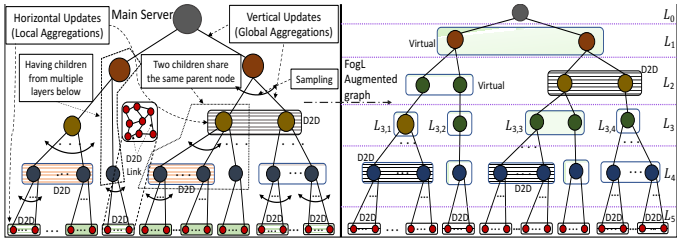


Fig. 3: Left: An example of FogL Network representation. The root corresponds to the main server, the leaves are the end devices, and the nodes in-between are intermediate fog nodes. In D2D-enabled clusters, the nodes form a certain topology over which the devices communicate with their neighbors for distributed model consensus. Right: The corresponding FogL augmented graph for analysis. Virtual nodes and clusters are highlighted in green.

II. System Model and Problem Formulation

In this section, we formalize the FogL network architecture (Sec. II-A) and the ML problem (Sec. II-B). Then, we formalize the hybrid learning paradigm via intra- and inter-cluster communications (Sec. II-C) and parameter sharing (Sec. II-D).

A. Network Architecture and Graph Model

We consider the network architecture of FogL as depicted on the left in Fig. 3. In FogL, both inter-layer and intra-layer communications take place to conduct ML model training. The inter-layer communications are captured via a tree graph, with the main server at the root and end devices as the leaves. Each layer is partitioned into multiple clusters, with each device in a cluster sharing the same parent node in the next layer up. In general, each node may have a parent node located multiple layers above (e.g., an edge device can be directly connected to an edge server), and multiple clusters can share the same parent node (e.g., multiple groups of cellular devices share the same BS). Except at the bottom layer, each node in the hierarchy is the *parent* for a subset of the nodes, and is responsible for gathering the model parameters of its *children* nodes.

From this FogL network representation, we construct the *FogL augmented network graph* shown on the right of Fig. 3, where several virtual nodes and clusters have been added. Virtual nodes are added in such a way that each node has a single parent node in its immediate upper layer. Also, when multiple clusters share the same parent node, a layer is added to the FogL augmented graph that consists of multiple intermediate virtual nodes forming a virtual cluster, such that there is always a one-to-one mapping between the clusters and parent nodes. If necessary, the nodes in the highest layer before the main server will form a virtual cluster to preserve this one-to-one mapping. A node without any neighbors in its layer is also assumed to form a virtual singleton cluster. For convenience, we refer to the structure of Fig. 3 as a *tree* since in macro-scale it resembles a tree structure. However, in the micro-scale, nodes inside the clusters form *connected* graphs through which D2D communications are performed, differentiating the structure from a tree graph.

The FogL augmented graph has the following properties: (i) each parent node has a single cluster associated with it in the layer immediately below it; (ii) the length of the paths from the root to each of the leaf nodes are the same; and (iii) the layer below the root always consists of one cluster. The network consists of $|\mathcal{L}|+1$ layers, where $\mathcal{L} \triangleq \{L_1, \cdots, L_{|\mathcal{L}|}\}$ denotes the set of layers below the server. The root/server is

located at L_0 and the end devices are contained in $L_{|\mathcal{L}|}$. Inside layer L_j, $1 \leq j \leq |\mathcal{L}|$, there exists a set of clusters indexed by $L_{j,1}, L_{j,2}, \cdots$ (see Fig. 3). We let the nodes move between the clusters in the same layer. To capture these dynamics, we use $\mathcal{L}_{j,i}^{(k)}$ (i.e., calligraphic font) to refer to the set of nodes inside cluster $L_{j,i}$ at learning iteration k (described in Sec. II-B). For ease of presentation, we assume that the number of clusters at each layer is time invariant, i.e., each cluster always contains at least one node and nodes do not form new clusters. We let $\mathcal{N}_j$ denote the set of nodes in layer L_j and $N_j \triangleq |\mathcal{N}_j|$.

For convenience, we will sometimes use C to denote an arbitrary cluster (i.e., any $L_{j,i}$) and $\mathcal{C}^{(k)}$ (i.e., calligraphic font) to refer to its set of nodes at iteration k. We also sometimes use n to denote an arbitrary node. For each node n located in layers $L_{|\mathcal{L}|-1}, \cdots, L_0$, we let $Q(n)$ denote the index of its child cluster and $\mathcal{Q}^{(k)}(n)$ denote the set of its children nodes (i.e., the nodes in $Q(n)$) at global iteration k.

B. Machine Learning Task

Each end device n is associated with a dataset $\mathcal{D}_n$. Each element $d_i \in \mathcal{D}_n$ of a dataset, called a training sample, is represented via a feature vector $\mathbf{x}_i$ and a label y_i for the ML task of interest. For example, in image classification, $\mathbf{x}_i$ may be the RGB colors of all pixels in the image, and y_i may be the identity of the person in the image sample. The goal of the ML task is to learn the parameters $\mathbf{w} \in \mathbb{R}^M$ of a particular M-dimensional model (e.g., an SVM or a neural network) that are expected to maximize the accuracy in mapping from $\mathbf{x}_i$ to y_i across any sample in the network. The model is associated with a loss $\widetilde{f}(\mathbf{w}, \mathbf{x}_i, y_i)$, referred to as $\widetilde{f}_i(\mathbf{w})$ for brevity, that quantifies the error of parameter realization $\mathbf{w}$ on d_i. We refer to Table 1 in [19] for a list of common ML loss functions.

The global loss of the ML model is formulated as

$$F(\mathbf{w}) = \frac{1}{D} \sum_{n \in \mathcal{N}_{|\mathcal{L}|}} |\mathcal{D}_n| f_n(\mathbf{w}), \quad D = \sum_{n \in \mathcal{N}_{|\mathcal{L}|}} |\mathcal{D}_n|, \tag{1}$$

where f_n is the local loss at node n, i.e., $f_n(\mathbf{w}) = \frac{1}{|\mathcal{D}_n|} \sum_{d_i \in \mathcal{D}_n} \widetilde{f}_i(\mathbf{w})$. The goal of model training is to identify the optimal parameter $\mathbf{w}^*$ that minimizes the global loss:

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{R}^M} F(\mathbf{w}). \tag{2}$$

To achieve this in a distributed manner, training is conducted through consecutive global iterations. At the start of global iteration $k \in \mathbb{N}$, the main server possesses a parameter vector $\mathbf{w}^{(k-1)} \in \mathbb{R}^M$, which propagates downstream through the hierarchy to the end devices. Each end device n overrides its current local model parameter vector $\mathbf{w}_n^{(k-1)}$ according to

$$\mathbf{w}_n^{(k-1)} = \mathbf{w}^{(k-1)}, \tag{3}$$

and then performs a local update using gradient descent [3] as

$$\mathbf{w}_n^{(k)} = \mathbf{w}_n^{(k-1)} - \beta \nabla f_n(\mathbf{w}_n^{(k-1)}), \ n \in \mathcal{N}_{|\mathcal{L}|}, \tag{4}$$

where β is the step-size. The main server wishes to obtain the global model used for initiating the next global iteration, which is defined as a weighted average of end devices' parameters:

$$\mathbf{w}^{(k)} = \frac{\sum_{n \in \mathcal{N}_{|\mathcal{L}|}} |\mathcal{D}_n| \mathbf{w}_n^{(k)}}{D}. \tag{5}$$

The value of D in (5) is assumed to be known at the main server, since it only requires uplink transmission of the scalar $|\mathcal{D}_n|$ by each end device $n \in \mathcal{N}_{|\mathcal{L}|}$, which can be aggregated and propagated upstream by each parent node in the hierarchy. As we will see, our method does not require knowledge of each individual $|\mathcal{D}_n|$ at the server to conduct the parameter averaging given in (5), since the number of data points at each end device will be encapsulated in a scaled model parameter vector shared to its parent node (see Sec. III-B). On the downlink from the server to the edge devices, we assume that the (common) global parameter $\mathbf{w}^{(k-1)}$ can be readily shared through the hierarchy to reach to the end devices, e.g., through a broadcasting protocol. These devices will then conduct (3) and (4) locally. The challenge, then, is computing (5) at the main server. To do this, in federated learning, the devices will directly upload (4) to the server. However, this is prohibitive in a large-scale fog computing system. First, it may require *high energy consumption*: uplink transmissions from battery-limited devices to nodes at a higher layer typically correspond to long physical distances, and can deplete individual device batteries. Second, it may induce *high network traffic and long latencies*: a neural network with even hundreds of parameters, which would be small by today's standards [24], would require transmission of billions of parameters in the upper layers during each iteration over a network with millions of nodes. Additionally, it may *overload current cellular and vehicular architectures*: these infrastructures are not designed to handle large jumps in the number of active users [43], which would be the case with simultaneous uplink transmissions at the bottom-most layer. These issues require a novel approach to parameter aggregations, which we develop in this paper.

C. Hybrid Learning via Intra- and Inter-Layer Communications

The main server in FogL is only interested in the weighted average of the local parameters (5). Consequently, we propose *local aggregations* at each network layer. To achieve this, each cluster in Fig. 3 follows one of two mechanisms:

1) *Distributed aggregation*: The nodes engage in a cooperative scheme facilitated by D2D communications to realize the consensus/average of their local model parameters. The parent node then samples parameters of one of the children and scales it by the number of children to calculate an approximate sum of the children nodes' parameters.
2) *Instant aggregation*: Each node instead uploads its local model to the parent node. The parent computes the aggregation directly as a sum of the children nodes' parameters.

The communication requirement of the instant aggregation is often significantly higher than the distributed aggregation since D2D communications generally occur over much shorter distances, which makes them less power/energy consuming. We refer to mechanisms 1 and 2 mentioned above as limited uplink transmission (LUT) and extensive uplink transmission (EUT), respectively.[1] Not all clusters can operate in LUT mode, since not all are D2D enabled, e.g., due to sparse connections between devices. Still, we can expect significant advantages in terms of the volume of parameters uploaded through the system

[1]We assume that each node belongs either to an EUT or to an LUT cluster. If some portion of the nodes in a cluster are capable of D2D communications while the rest are not, the cluster can be broken down into two clusters (i.e., an LUT and an EUT cluster) with the same parent node, based on which the fogL augmented network graph in Sec. II-A is then constructed.

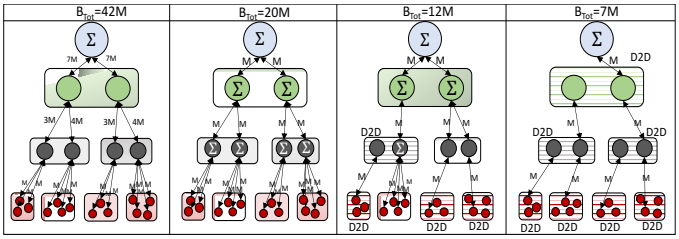


Fig. 4: Example of the network traffic reduction provided by multi-layer aggregations in FogL, where each device trains a model with parameter length M. The sum of the length of parameters transferred among the layers (B_{Tot}) is depicted at the top. (a) Network consisting of all EUT clusters, where the parent nodes upload all received parameters from their children. (b) Network consisting of all EUT clusters, where the parent nodes sum all received parameters and upload to the next layer. (c) Network with a portion of clusters in LUT mode, where each parent node only samples the parameters of one device after consensus formation. (d) Network consisting of all LUT clusters.

with a combination of EUT and LUT clusters, as depicted in Fig. 4. At the bottom layers where communication is mostly over the air, D2D can also be implemented through the out-band mode [44]. This has the additional advantage of not occupying the licensed spectrum, which results in bandwidth savings that can lead to an improved quality of service.

We allow a cluster to switch between EUT and LUT over time. For instance, the connectivity between a fleet of miniature UAVs will vary as they travel, necessitating EUT when D2D is not feasible. To capture this dynamic, for each cluster C, at global iteration k, $\mathbb{1}_{\{C\}}^{(k)}$ captures the operating mode, which is 1 if the cluster operates in LUT mode, and 0 otherwise. In each LUT cluster, a node will only communicate with its neighboring devices, which may not include the whole cluster. Further, each node's neighborhood may evolve over time. For example, when the communications are conducted over the air as in Fig. 2(a), the neighbors in one aggregation interval are identified based on the distances among the nodes and their transmit powers. We will explicitly consider such evolutions in cluster topology in our distributed consensus model in Sec. III.

In the simple case where there is only one layer below the server, i.e., $|\mathcal{L}| = 1$, consisting of a single EUT cluster, the FogL architecture reduces to federated learning. If instead there is just one layer of one LUT cluster with no server, FogL resembles fully distributed learning [17], [28], [29]. One of our contributions is developing and analyzing this generalized cluster-based multi-layer hybrid learning paradigm for FogL.

D. Parameter Sharing vs. Gradient Sharing

Note that the parameter update in (5) can be written as

$$\begin{aligned}\mathbf{w}^{(k)} &= \frac{\sum_{n\in\mathcal{N}_{|\mathcal{L}|}} |\mathcal{D}_n| \left(\mathbf{w}_n^{(k-1)} - \beta\left(\nabla f_n(\mathbf{w}_n^{(k-1)})\right)\right)}{D} \\ &= \mathbf{w}^{(k-1)} - \beta\frac{\sum_{n\in\mathcal{N}_{|\mathcal{L}|}} |\mathcal{D}_n| \nabla f_n(\mathbf{w}_n^{(k-1)})}{D}.\end{aligned} \tag{6}$$

This asserts that the global parameters can also be obtained via the gradients of the devices, implying that the devices can either share their gradients or their parameters during training. This equivalence arises from the one-step update in (4), which is a common assumption in federated learning [13], [14], [45]. However, recent work [19], [21] has advocated conducting multiple rounds of local updates between global aggregations to reduce communication costs; in this case, the parameters are required for each aggregation. In this paper, we focus on the more general case of parameter sharing, although we obtain one theoretical result (Proposition 3) based on gradient sharing.

In LUT clusters, devices leverage D2D communications to obtain an approximate value of the average of their parameters. A basic approach would be to implement a message passing algorithm where nodes exchange parameters with their neighbors until each node in the cluster has all parameters stored locally. Each node can then readily calculate the aggregated value, and one of them can be sampled by the parent node. Collecting a table of parameters at each node may not be feasible, however, given how large these vectors can be for contemporary ML models, as discussed in Sec. II-B. Instead, we desire a technique that (i) does not require any node to store a table of all model parameters in the cluster, (ii) can be implemented in a distributed manner via D2D, and (iii) is generalizable across different network layers. In the following, we leverage *distributed average consensus* methods for this.

III. Multi-Stage Hybrid Federated Learning(MH−FL)

In this section, we develop our MH−FL methodology (Sec. III-A&III-B). Then, we conduct a detailed performance analysis of our method (Sec. III-C). Finally, based on this analysis, we develop online control algorithms for tuning the number of D2D rounds in each cluster over time to guarantee the convergence properties for our method (Sec. III-D).

A. Distributed Average Consensus within Clusters

Referring to Fig. 3, during each iteration k of training, the LUT clusters engage in D2D communications, where each node desires estimating the average value of the parameters inside its cluster. For each cluster C that is LUT during the kth global aggregation iteration (i.e., for which $\mathbb{1}_{\{C\}}^{(k)} = 1$), we let $G_C^{(k)} = \left(\mathcal{C}^{(k)}, \mathcal{E}_C^{(k)}\right)$ denote its communication graph. In this graph, $\mathcal{C}^{(k)}$ is the set of nodes belonging to the cluster, and there is an edge $(m,n) \in \mathcal{E}_C^{(k)}$ between nodes $m, n \in \mathcal{C}^{(k)}$ iff they can communicate via D2D during k. We assume that $G_C^{(k)}$ is undirected, connected, and static for the duration of one iteration k, although it may vary in different iterations. Note that we have abstracted the physical layer wireless/wired communication medium of each LUT cluster C to a general graph topology $G_C^{(k)}$. In a wireless cluster, different channel conditions and communication configurations among the devices will manifest as different topologies for $G_C^{(k)}$.

We employ the family of *linear distributed consensus* algorithms [46], where during the global iteration k, the nodes inside LUT cluster C conduct $\theta_C^{(k)} \in \mathbb{N}$ *rounds* of D2D (number of D2D rounds is a design parameter obtained in Sec. III-C). Each round of D2D consists of parameter transfers between neighboring nodes. Formally, during global iteration k, each node $n \in \mathcal{C}^{(k)}$ engages in the following rounds of iterative updates for $t = 0, ..., \theta_C^{(k)} - 1$:

$$\mathbf{z}_n^{(t+1)} = v_{n,n}^{(k)} \mathbf{z}_n^{(t)} + \sum_{m\in\zeta^{(k)}(n)} v_{n,m}^{(k)} \mathbf{z}_m^{(t)}, \tag{9}$$

where $\mathbf{z}_n^{(0)} = \mathbf{w}_n^{(k)}$ corresponds to node n's initial parameter, and $\mathbf{z}_n^{\left(\theta_C^{(k)}\right)}$ denotes the parameter after the D2D consensus

$$\widehat{\mathbf{w}}_{n'}^{(k)} = \frac{\sum_{j\in\mathcal{Q}^{(k)}(n)} |\mathcal{D}_j|\mathbf{w}_j^{(k-1)}}{|\mathcal{Q}^{(k)}(n)|} - \frac{\sum_{j\in\mathcal{Q}^{(k)}(n)} \beta|\mathcal{D}_j|\nabla f_j(\mathbf{w}_j^{(k-1)})}{|\mathcal{Q}^{(k)}(n)|} + \mathbb{1}_{\{Q(n)\}}^{(k)}\mathbf{c}_{n'}^{(k)},\ n'\in\mathcal{Q}^{(k)}(n) \tag{7}$$

$$\widehat{\mathbf{w}}_{n'}^{(k)} = \frac{\sum_{i\in\mathcal{Q}^{(k)}(n)}\sum_{j\in\mathcal{Q}^{(k)}(i)} |\mathcal{D}_j|\mathbf{w}_j^{(k-1)}}{|\mathcal{Q}^{(k)}(n)|} - \frac{\sum_{i\in\mathcal{Q}^{(k)}(n)}\sum_{j\in\mathcal{Q}^{(k)}(i)} \beta|\mathcal{D}_j|\nabla f_j(\mathbf{w}_j^{(k-1)})}{|\mathcal{Q}^{(k)}(n)|} + \frac{\sum_{i\in\mathcal{Q}^{(k)}(n)} \mathbb{1}_{\{Q(i)\}}^{(k)}|\mathcal{Q}^{(k)}(i)|\mathbf{c}_{i'}^{(k)}}{|\mathcal{Q}^{(k)}(n)|} + \mathbb{1}_{\{Q(n)\}}^{(k)}\mathbf{c}_{n'}^{(k)}, n'\in\mathcal{Q}^{(k)}(n) \tag{8}$$

process. $\zeta^{(k)}(n)$ denotes the set of neighbors of node n during global iteration k, and $v_{n,p}^{(k)}$, $p\in\{n\}\cup\zeta^{(k)}(n)$ are the *consensus weights* associated with node n during k. There are several potential choices for these weights that guarantee convergence of the distributed consensus iterations, provided that the cluster graph $G_C^{(k)}$ is connected [46] (if it is not, we can partition the cluster into multiple connected subgraphs with the same parent node). We will detail the conditions required for convergence of local model aggregations in Assumption 2 of Sec. III-C. One choice that satisfies the conditions is $\mathbf{z}_n^{(t+1)} = \mathbf{z}_n^{(t)} + d_C^{(k)}\sum_{m\in\zeta^{(k)}(n)}(\mathbf{z}_m^{(t)} - \mathbf{z}_n^{(t)})$, $0 < d_C^{(k)} < 1/D_C^{(k)}$, where $D_C^{(k)}$ is the maximum degree of the nodes in $G_C^{(k)}$ [46]. With this constant edge weight implementation, the nodes inside LUT cluster C only need to have knowledge of the parameter $d_C^{(k)}$ to conduct local aggregations, which can be broadcast by the respective parent node. This choice of consensus weights is used in our simulations in Sec. IV.

Due to time constraints (i.e., depending on the required time between global aggregations), the number of D2D rounds cannot be arbitrarily large, and thus the nodes inside a cluster often do not have a perfect estimate of the average value of their parameters. In Sec. III-C, we analyze the effect of a finite number of D2D rounds on the `MH-FL` performance.

B. Local Aggregations and Parameters Propagation

In the following, we introduce a *scaled parameter* for each node n denoted by $\widetilde{\mathbf{w}}_n$ and develop an approach to perform global aggregations based on manipulation and relaying of these parameters across different layers of the network. The definition of $\widetilde{\mathbf{w}}_n$ depends on the layer where node n is located.

1) Nodes' parameters in the bottom-most layer: Given $\mathbf{w}^{(k-1)}$, the end devices in layer $L_{|\mathcal{L}|}$ first perform the local update described in (4). Since the nodes' parameters go through multiple stages of aggregations (see Fig. 4), the server cannot recover the individual parameters from the aggregated ones to calculate (5). To overcome this issue, each device $n\in\mathcal{N}_{|\mathcal{L}|}$ obtains its scaled parameter as $\widetilde{\mathbf{w}}_n^{(k)} = |\mathcal{D}_n|\mathbf{w}_n^{(k)}$ and shares it with its neighbors during the D2D process for global iteration k, i.e., its parameters weighted by its number of datapoints.[2] Using this weighting technique, the number of datapoints of the nodes is encoded in the multi-layer aggregations. Specifically, each node $n\in\mathcal{C}^{(k)}$ belonging to cluster C located in $L_{|\mathcal{L}|}$ engages in the local iterations described by (9), where $\mathbf{z}_n^{(0)} = \widetilde{\mathbf{w}}_n^{(k)}$. Finally, the node stores $\widehat{\mathbf{w}}_n^{(k)} = \mathbf{z}_n^{\left(\theta_C^{(k)}\right)}$, which corresponds to the final weighted local parameter value after the D2D process.

2) Nodes' parameters in the middle layers: Once D2D communications are finished in layer $L_{|\mathcal{L}|}$, each parent node $n\in\mathcal{N}_{|\mathcal{L}|-1}$ of a cluster that operated in LUT mode selects a cluster head n' among its children $\mathcal{Q}^{(k)}(n)$ in layer $L_{|\mathcal{L}|}$ based on a selection/sampling distribution. This child n' uploads its parameter vector $\widehat{\mathbf{w}}_{n'}^{(k)}$ to the parent node. The resulting sampled parameter vector at the parent node is given by (7), where the first two terms correspond to the true average of the parameters of the nodes inside cluster $Q(n)$, and $\mathbf{c}_{n'}^{(k)}\in\mathbb{R}^M$ is the error arising from the consensus. This error is only applicable to the LUT clusters and is concerned with the deviation from the true cluster average (which would be obtained from an EUT cluster). The parent node $n\in\mathcal{N}_{|\mathcal{L}|-1}$ then computes its scaled parameter $\widetilde{\mathbf{w}}_n^{(k)}$ by scaling the received vector by the number of its children, and stores the corresponding vector:

$$\widetilde{\mathbf{w}}_n^{(k)} = |\mathcal{Q}^{(k)}(n)|\widehat{\mathbf{w}}_{n'}^{(k)}, n\in\mathcal{N}_{|\mathcal{L}|-1}, n'\in\mathcal{Q}^{(k)}(n), \mathbb{1}_{\{Q(n)\}}^{(k)} = 1. \tag{10}$$

Also, in layer $\mathcal{L}_{|\mathcal{L}|-1}$, each parent node n of a cluster that operated in EUT mode receives all the parameters of its children and computes $\widetilde{\mathbf{w}}_n^{(k)} = \sum_{i\in\mathcal{Q}^{(k)}(n)}\widetilde{\mathbf{w}}_i^{(k)}$. Once this is completed, the LUT clusters located in layer $\mathcal{L}_{|\mathcal{L}|-1}$ engage in distributed consensus formation via cooperative D2D.

This procedure continues up the hierarchy at each layer L_j, $j = |\mathcal{L}|-2, ..., 1$. More precisely, for an LUT cluster C located in one of the middle layers, each node $n\in\mathcal{C}^{(k)}$ performs the local iterations described by (9) with initialization $\mathbf{z}_n^{(0)} = \widetilde{\mathbf{w}}_n^{(k)}$. At the end of the D2D rounds, each of these nodes stores the last obtained parameter $\widehat{\mathbf{w}}_n^{(k)} = \mathbf{z}_n^{\left(\theta_C^{(k)}\right)}$, and passes it up if sampled by its parent. At the same time, each node n inside an EUT cluster located in one of the middle layers directly shares $\widetilde{\mathbf{w}}_n^{(k)}$ for the calculation of the local aggregation.

Traversing up the layers, the consensus errors accumulate. For example, (8) gives the expression for the final consensus parameter vector of a node $n'\in\mathcal{Q}^{(k)}(n)$ inside an LUT cluster at layer $L_{|\mathcal{L}|-1}$, which will be sampled by a parent node $n\in\mathcal{N}_{|\mathcal{L}|-2}$ located in the upper layer. In this expression, i' denotes the sampled node chosen by parent node i.

3) Main server: Let $\widetilde{\mathbf{w}}_0^{(k)}$ denote the scaled parameter at the main server. If the cluster in layer L_1 operates in LUT mode, then $\widetilde{\mathbf{w}}_0^{(k)} = |\mathcal{L}_{1,1}^{(k)}|\widehat{\mathbf{w}}_m^{(k)}$, where $m\in\mathcal{L}_{1,1}^{(k)}$ denotes the node sampled by the main server with parameter $\widehat{\mathbf{w}}_m^{(k)}$ obtained after performing D2D rounds[3]. Otherwise, the main server sums all the received parameters, i.e., $\widetilde{\mathbf{w}}_0^{(k)} = \sum_{m\in\mathcal{L}_{1,1}^{(k)}}\widetilde{\mathbf{w}}_m^{(k)}$. The main server then computes the global parameter vector as

$$\mathbf{w}^{(k)} = \widetilde{\mathbf{w}}_0^{(k)}/D, \tag{11}$$

which is then broadcast down the hierarchy to start the next global round $k+1$, beginning with local updates at the devices.

[2]Nodes inside EUT clusters of layer $L_{|\mathcal{L}|}$ directly share their scaled parameters with their parents.

[3]According to FogL augmented graph properties, L_1 consists of one cluster, and thus $|\mathcal{L}_{1,1}^{(k)}| = |\mathcal{N}_1|$ which is inherently time invariant. The super-index k is added for consistency in calligraphic notations.

Algorithm 1: Multi-stage hybrid federated learning(MH−FL)

input : number of global aggregations K, default number of D2D rounds $\theta^{(k)}_{L_{j,i}}\ \forall i,j,k$.

output : Final global model $\mathbf{w}^{(K)}$.

Initial operations at main server: Initialize the global parameter $\mathbf{w}^{(0)}$ and synchronize the edge devices with it.

** **Initial operations at main server:** If asymptotic convergence to optimal is desired: (i) server randomly sets $\|\nabla\widehat{F(\mathbf{w}^0)}\|$ and broadcasts it, (ii) Server sets δ either arbitrarily or according to (24) to guarantee a certain accuracy, and broadcasts it. Otherwise, the server sets D2D control parameters $\{\sigma_j\}_{j=1}^{|\mathcal{L}|}$ as described in Sec. III-D and broadcasts them.

for $k=1$ *to* K **do**

  for $l=|\mathcal{L}|$ *down to* $l=0$ **do**

    if $l=|\mathcal{L}|$ **then**

      Given $\mathbf{w}^{(k-1)}$, each node n obtains $\mathbf{w}_n^{(k)}$ using (4).

      Each node n obtains its scaled parameter $\widetilde{\mathbf{w}}_n^{(k)}=|\mathcal{D}_n|\mathbf{w}_n^{(k)}$.

      ** Each cluster operating in LUT mode runs Algorithm 2.

      ## Each cluster operating in LUT mode runs Algorithm 3.

      Nodes inside LUT clusters update their parameters using (9).

    else if $1\leq l\leq|\mathcal{L}|-1$ **then**

      Each parent node n of an LUT cluster $Q(n)$ samples a child $n'\in\mathcal{Q}^{(k)}(n)$ and computes $\widetilde{\mathbf{w}}_n^{(k)}=|\mathcal{Q}^{(k)}(n)|\widehat{\mathbf{w}}_{n'}^{(k)}$.

      Each parent node n of an LUT cluster $Q(n)$ uses the received parameters to compute $\widetilde{\mathbf{w}}_n^{(k)}=\sum_{i\in\mathcal{Q}^{(k)}(n)}\widetilde{\mathbf{w}}_i^{(k)}$.

      ** Each cluster operating in LUT mode runs Algorithm 2.

      ## Each cluster operating in LUT mode runs Algorithm 3.

      Nodes inside LUT clusters update their parameters using (9).

    else if $l=0$ **then**

      if $\mathbb{1}^{(k)}_{\{L_{1,1}\}}=1$ **then**

        The server computes $\widetilde{\mathbf{w}}_0^{(k)}=|\mathcal{L}_{1,1}^{(k)}|\widehat{\mathbf{w}}_m^{(k)}$, where $m\in\mathcal{L}_{1,1}^{(k)}$.

      else

        The server computes $\widetilde{\mathbf{w}}_0^{(k)}=\sum_{m\in\mathcal{L}_{1,1}^{(k)}}\widetilde{\mathbf{w}}_m^{(k)}$.

      The server computes $\mathbf{w}^{(k)}$ using (11) and broadcasts it.

      ** If asymptotic convergence to optimal is desired, the main server approximates the gradient of the loss function used for the next iteration as in Sec. III-D and broadcasts it.

The MH−FL methodology we developed throughout this section is summarized in Algorithm 1. The lines beginning with ** and ## are enhancements for tuning the D2D rounds over time at different clusters, which we present in Sec. III-D.

C. Theoretical Analysis of MH−FL

One of the key contributions of MH−FL is the integration of D2D communications with multi-layer parameter transfers. As discussed, D2D communications are conducted through time varying topology structures among the nodes, which introduce a network dimension to model training. Studying this effect under limited D2D rounds regime is the main theme of our theoretical analysis. Before presenting our main results, we first introduce a few assumptions and a definition. Henceforth, $\|.\|$ denotes the 2-norm unless otherwise stated.

Assumption 1. *The global ML loss function* (1) *has the following properties: (i)* μ*-strong convexity, i.e.,* $F(\mathbf{y})\geq F(\mathbf{x})+(\mathbf{y}-\mathbf{x})^\top\nabla F(\mathbf{x})+\frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|^2$ *for some* $\mu>0,\ \forall\mathbf{x},\mathbf{y}$*, and (ii)* η*-smoothness, i.e,* $\|\nabla F(\mathbf{x})-\nabla F(\mathbf{y})\|\leq\eta\|\mathbf{x}-\mathbf{y}\|$ *for some* $\eta>\mu,\ \forall\mathbf{x},\mathbf{y}$.

The above properties are common assumptions in federated learning and ML literature [13], [45], [47]–[49]. Commonly encountered ML models with convex loss functions are linear regression, logistic regression, squared SVM, and single layer neural networks with convex activation functions. In practice, these models are implemented with an additional regularization term to improve the convergence and avoid model overfitting, which makes them strongly convex [49]. We will conduct our convergence analysis based on this assumption and design control algorithms in Sec. III-D for both the convex (Algorithm 2) and non-convex (Algorithm 3) cases.

We will also find it useful to write the consensus algorithm in matrix form. Letting $\widetilde{\mathbf{W}}_C^{(k)}\in\mathbb{R}^{|\mathcal{C}^{(k)}|\times M}$ denote the matrix of scaled parameters across all nodes in an LUT cluster C prior to consensus in iteration k, the evolution of the nodes' parameters described by (9) can be written as

$$\widehat{\mathbf{W}}_C^{(k)}=\left(\mathbf{V}_C^{(k)}\right)^{\theta_C^{(k)}}\widetilde{\mathbf{W}}_C^{(k)},\tag{12}$$

where $\widehat{\mathbf{W}}_C^{(k)}\in\mathbb{R}^{|\mathcal{C}^{(k)}|\times M}$ denotes the matrix of node parameters after the consensus, and $\mathbf{V}_C^{(k)}=[v_{n,m}^{(k)}]_{n,m\in\mathcal{C}^{(k)}}$ is the consensus matrix applied to the parameter vector to realize (9).

Assumption 2. *The consensus matrix* $\mathbf{V}_C^{(k)}$ *for each LUT cluster* C *has the following properties [46], [50]: (i)* $\left(\mathbf{V}_C^{(k)}\right)_{m,n}=0$ *if* $(m,n)\notin\mathcal{E}_C^{(k)}$*, (ii)* $\mathbf{V}_C^{(k)}\boldsymbol{1}=\boldsymbol{1}$*, (iii)* $\mathbf{V}_C^{(k)}={\mathbf{V}_C^{(k)}}^\top$*, and (iv)* $\rho\left(\mathbf{V}_C^{(k)}-\frac{\boldsymbol{1}\boldsymbol{1}^\top}{|\mathcal{C}^{(k)}|}\right)\leq\lambda_C^{(k)}<1$*, where* $\boldsymbol{1}$ *is the vector of 1s and* $\rho(\mathbf{A})$ *is the spectral radius of matrix* $\mathbf{A}$.

In Assumption 2, $\lambda_C^{(k)}$ can be interpreted as an upper bound on the spectral radius, which plays a key role in our results.

Definition 1. *The divergence of parameters in cluster* C *at iteration* k*, denoted by* $\Upsilon_C^{(k)}$*, is defined as an upper bound on the difference of its nodes' scaled parameters as follows:*

$$\left\|\widetilde{\mathbf{w}}_q^{(k)}-\widetilde{\mathbf{w}}_{q'}^{(k)}\right\|\leq\Upsilon_C^{(k)},\ \forall q,q'\in\mathcal{C}^{(k)}.\tag{13}$$

At the bottom-most layer, the above defined quantity is indicative of the degree of data heterogeneity (i.e., the level of non-i.i.d) among the nodes in a cluster, whereas in the upper layers it captures the heterogeneity of data contained in sub-trees with their roots being the nodes in the cluster. We show in Theorem 1 how it impacts the convergence bound, and in the subsequent results how it dictates the number of D2D rounds. Then, in Sec. III-D, we develop control algorithms that approximate the divergence in a distributed manner at every cluster, and use it to control the convergence rate.

1) General convergence bound: In the following theorem, we study the convergence of MH−FL (see Appendix A):

Theorem 1. *With a learning rate* $\beta=1/\eta$*, after* k *global iterations of any realization of* MH−FL*, an upper bound on* $F(\mathbf{w}^{(k)})-F(\mathbf{w}^\star)$ *is given in* (14)*, where* $\Phi=N_{|\mathcal{L}|-1}+N_{|\mathcal{L}|-2}+\cdots+N_1+1$ *is the total number of nodes in the network besides the bottom layer and* $\Xi^{(k-t)}$ *is given by* (15)*.*

Remark 1. *Each nested sum in* (15)*, e.g., (b), encompasses the nodes located in a path starting from the main server and ending at a node in one of the layers. Different nested sums capture paths with different lengths. Also, each summand, e.g., (c), corresponds to the characteristics of the child cluster, e.g.,* $Q(a_{|\mathcal{L}|-1})$ *in (c), of the node in the last index of its associated nested sum, e.g.,* $a_{|\mathcal{L}|-1}$ *in (b). This contains the operating mode, number of nodes, upper bound of spectral radius, number of D2D rounds, and divergence of parameters.*

Main takeaways. The bound in (14), (15) quantifies how the convergence is dependent on several learning and system parameters. In particular, we see a dependence on (i) the charac-

$$F(\mathbf{w}^{(k)}) - F(\mathbf{w}^*) \leq \underbrace{\left(1-\frac{\mu}{\eta}\right)^k \left(F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*)\right)}_{\text{(a)}} + \frac{\eta\Phi}{2D^2} \sum_{t=0}^{k-1} \left(1-\frac{\mu}{\eta}\right)^t \Xi^{(k-t)} \tag{14}$$

$$\begin{aligned}\Xi^{(k-t)} = & \underbrace{\sum_{a_1 \in \mathcal{L}_{1,1}^{(k-t)}} \sum_{a_2 \in \mathcal{Q}^{(k-t)}(a_1)} \cdots \sum_{a_{|\mathcal{L}|-1} \in \mathcal{Q}^{(k-t)}(a_{|\mathcal{L}|-2})}}_{(b)} \underbrace{\mathbb{1}^{(k-t)}_{\left\{Q(a_{|\mathcal{L}|-1})\right\}} |\mathcal{Q}^{(k-t)}(a_{|\mathcal{L}|-1})|^3 \left(\lambda^{(k-t)}_{Q(a_{|\mathcal{L}|-1})}\right)^{2\theta^{(k-t)}_{Q(a_{|\mathcal{L}|-1})}} \left(\Upsilon^{(k-t)}_{Q(a_{|\mathcal{L}|-1})}\right)^2}_{(c)} \\ & + \sum_{a_1 \in \mathcal{L}_{1,1}^{(k-t)}} \sum_{a_2 \in \mathcal{Q}^{(k-t)}(a_1)} \cdots \sum_{a_{|\mathcal{L}|-2} \in \mathcal{Q}^{(k-t)}(a_{|\mathcal{L}|-3})} \mathbb{1}^{(k-t)}_{\left\{Q(a_{|\mathcal{L}|-2})\right\}} |\mathcal{Q}^{(k-t)}(a_{|\mathcal{L}|-2})|^3 \left(\lambda^{(k-t)}_{Q(a_{|\mathcal{L}|-2})}\right)^{2\theta^{(k-t)}_{Q(a_{|\mathcal{L}|-2})}} \left(\Upsilon^{(k-t)}_{Q(a_{|\mathcal{L}|-2})}\right)^2 \\ & + \cdots + \sum_{a_1 \in \mathcal{L}_{1,1}^{(k-t)}} \mathbb{1}^{(k-t)}_{\{Q(a_1)\}} |\mathcal{Q}^{(k-t)}(a_1)|^3 \left(\lambda^{(k-t)}_{Q(a_1)}\right)^{2\theta^{(k-t)}_{Q(a_1)}} \left(\Upsilon^{(k-t)}_{Q(a_1)}\right)^2 + \mathbb{1}^{(k-t)}_{\left\{L_{1,1}\right\}} |\mathcal{L}^{(k-t)}_{1,1}|^3 \left(\lambda^{(k-t)}_{L_{1,1}}\right)^{2\theta^{(k-t)}_{\mathcal{L}_{1,1}}} \left(\Upsilon^{(k-t)}_{L_{1,1}}\right)^2\end{aligned} \tag{15}$$

teristics of the loss function (i.e., η, μ), (ii) the number of nodes and clusters at each network layer (through the $|\mathcal{Q}|$ terms), (iii) the topology and characteristics of the communication graph among the nodes inside the clusters, captured via the spectral radius bounds (i.e., λ), (iv) the number of D2D rounds performed at each cluster (i.e., the θ), and (v) the divergence among the node parameters at each cluster (i.e., Υ). Given a fixed set of parameters at iteration $k-1$, we can observe that increasing the number of D2D rounds at each cluster in iteration k results in a smaller bound (since θ is appeared as the exponent of $\lambda < 1$), and thus a better expected model loss, as we would expect. Furthermore, for a fixed number of D2D rounds, a smaller spectral radius, corresponding to a better connected cluster, results in a smaller bound. On the other hand, larger parameter divergence results in a worse bound.

Term (a) in (14) corresponds to the case with no consensus error in the system, i.e., when all LUT clusters have $\theta_C^{(k)} \to \infty$ (infinite D2D rounds) or when the network consists of all EUT clusters. Since $1-\mu/\eta < 1$, the overall rate of convergence of MH-FL is at best linear with rate $1-\mu/\eta$. However, achieving this would incur prohibitively long delays, motivating us to study the effects of the number of D2D rounds. Also, note that the terms $1-\left(\mu/\eta\right)^t$ inside the summation have a dampening effect: at global iteration k, the errors from global iteration $t < k$ are multiplied by $1-\left(\mu/\eta\right)^{k-t}$, meaning the initial errors for $t \ll k$ are dampened by very small coefficients, while the final errors have a more pronounced effect on the bound. At first glance, this seems to suggest that at higher global iteration indices, more D2D rounds are needed to reduce the errors. However, especially upon having i.i.d datasets, we can expect the parameters of the end devices to become more similar to one another with increasing global iteration count, which in turn would decrease the divergence (i.e., Υ) within clusters over time. Further, the connectivity of the clusters (captured by λ) will change at different layers of the hierarchy, causing the spectral radius to vary. This motivates us in Sec. III-D to consider adapting the D2D rounds over two dimensions: time (i.e., global iterations) and space (i.e., network layers).

2) Asymptotic optimality: We now explicitly connect the number of D2D rounds performed at different clusters with the asymptotic optimality of MH-FL (see Appendix B).

Proposition 1. *For any realization of* MH-FL*, if the number of D2D rounds at different clusters in different layers of the network satisfies the following criterion ($\forall k, i, j$):*

$$\begin{cases} \theta^{(k)}_{L_{j,i}} \geq \left\lceil \frac{\log\left(\sigma_j\right) - 2\log\left(\left|\mathcal{L}^{(k)}_{j,i}\right|^{\frac{3}{2}} \Upsilon^{(k)}_{L_{j,i}}\right)}{2\log\left(\lambda^{(k)}_{L_{j,i}}\right)} \right\rceil, & \text{if } \sigma_j \leq \left|\mathcal{L}^{(k)}_{j,i}\right|^3 \left(\Upsilon^{(k)}_{L_{j,i}}\right)^2 \\ \theta^{(k)}_{L_{j,i}} \geq 0, & \text{otherwise} \end{cases} \tag{16}$$

for non-negative constants $\sigma_1, \cdots, \sigma_{|\mathcal{L}|}$, then the asymptotic upper bound on the distance from the optimal is given by

$$\lim_{k\to\infty} F(\mathbf{w}^{(k)}) - F(\mathbf{w}^*) \leq \frac{\eta^2\Phi}{2\mu D^2} \sum_{j=0}^{|\mathcal{L}|-1} \sigma_{j+1} N_j. \tag{17}$$

Proposition 1 gives a guideline for designing the number of D2D rounds at different network clusters over time to achieve a desired (finite) upper bound on the optimality gap. It can be seen that optimality is tied to our introduced auxiliary variables $\{\sigma_j\}_{j=1}^{|\mathcal{L}|}$, which we refer to as *D2D control parameters*.

To eliminate the existence of a constant optimality gap in (17), we obtain an extra condition on the tuning of these parameters and make them time-varying to guarantee a linear convergence to the optimal solution (see Appendix C):

Proposition 2. *For any realization of* MH-FL*, suppose that the number of D2D rounds at different clusters of different network layers satisfies the following criterion ($\forall k, i, j$):*

$$\begin{cases} \theta^{(k)}_{L_{j,i}} \geq \left\lceil \frac{\log\left(\sigma^{(k)}_j\right) - 2\log\left(\left|\mathcal{L}^{(k)}_{j,i}\right|^{\frac{3}{2}} \Upsilon^{(k)}_{L_{j,i}}\right)}{2\log\left(\lambda^{(k)}_{L_{j,i}}\right)} \right\rceil, & \text{if } \sigma^{(k)}_j \leq \left|\mathcal{L}^{(k)}_{j,i}\right|^3 \left(\Upsilon^{(k)}_{L_{j,i}}\right)^2 \\ \theta^{(k)}_{L_{j,i}} \geq 0, & \text{otherwise} \end{cases} \tag{18}$$

where the non-negative constants $\sigma^{(k)}_1, \cdots, \sigma^{(k)}_{|\mathcal{L}|}$ satisfy

$$\sum_{j=0}^{|\mathcal{L}|-1} \sigma^{(k)}_{j+1} N_j \leq \frac{D^2\mu(\mu-\delta\eta)}{\eta^4\Phi} \left\|\nabla F(\mathbf{w}^{(k-1)})\right\|^2 \tag{19}$$

for $0 < \delta \leq \mu/\eta$. Then, we have

$$F(\mathbf{w}^{(k+1)}) - F(\mathbf{w}^*) \leq (1-\delta)\Big(F(\mathbf{w}^{(k)}) - F(\mathbf{w}^*)\Big), \ \forall k, \tag{20}$$

which implies a linear convergence of MH-FL *and* $\lim_{k\to\infty} F(\mathbf{w}^{(k)}) - F(\mathbf{w}^*) = 0$.

Proposition 2 asserts that under a stricter tuning of the number of D2D rounds at different layers, i.e., (18) and (19), convergence to the optimal solution can be guaranteed with a rate that is at most $1-\mu/\eta$ according to (20). Furthermore, the number of D2D rounds are always finite when all the D2D tuning variables are greater than zero; according to (19), this can be always satisfied until reaching the optimal point (the

$$\kappa \geq \left\lceil \frac{\log\left(\epsilon - \frac{\eta^2\Phi}{2\mu D^2}\sum_{j=0}^{|\mathcal{L}|-1}\sigma_{j+1}N_j\right) - \log\left(F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*) - \frac{\eta^2\Phi}{2\mu D^2}\sum_{j=0}^{|\mathcal{L}|-1}\sigma_{j+1}N_j\right)}{\log\left(1-\mu/\eta\right)} \right\rceil \quad (21) \qquad \kappa \geq \left\lceil \frac{\log(\epsilon) - \log(F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*))}{\log(1-\delta)} \right\rceil \quad (22)$$

gradient of loss becomes zero) where the algorithm will stop. Note that Proposition 2's condition boosts the required $\theta^{(k)}_{L_{j,i}}$ over time, since the norm of the gradient in (19) decreases over time, in turn lowering the values of $\{\sigma_j^{(k)}\}_{j=1}^{|\mathcal{L}|}$. In Proposition 1, by contrast, the D2D rounds will become *tapered* over time (i.e, they will diminish over time), since σ_j is fixed over k and the divergence of the parameters is expected to decrease over global iterations, especially when dealing with i.i.d data. Our experiments in Sec. IV verify these effects.

Proposition 2's result also assumes knowledge of the global loss gradient $\left\|\nabla F(\mathbf{w}^{(k-1)})\right\|$, which is not known at the beginning of global iteration k, where $\mathbf{w}^{(k-1)}$ is just sent down through the layers. In Sec. III-D, we will develop an approximation technique for implementing this result in practice. Finally, note that in both Propositions 1&2, a smaller spectral radius (corresponding to well connected clusters) is tied to a lower number of D2D rounds (note that $\lambda < 1$).

3) Relationship between global iterations and D2D rounds: The following two corollaries to Propositions 1&2 investigate the impact of the number of global iterations on the required D2D rounds, and vice versa. First, we obtain the number of D2D rounds required at different clusters to reach a desired accuracy in a desired global iteration (see Appendix D):

Corollary 1. *Let $\epsilon \in \left[(1-\frac{\mu}{\eta})^{\kappa}\left(F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*)\right),\ F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*)\right)$. To guarantee that* MH−FL *obtains a solution to within ϵ of the optimal by global iteration κ, i.e., $F(\mathbf{w}^{(\kappa)}) - F(\mathbf{w}^*) \leq \epsilon$, a sufficient number of D2D rounds in different clusters of the network is given by either of the following conditions:*

1) *$\theta^{(k)}_{L_{j,i}}$, $\forall i,j,k$, given by* (16)*, where the values of $\sigma_1, \cdots, \sigma_{|\mathcal{L}|}$ satisfy the following inequality:*
$$\sum_{j=0}^{|\mathcal{L}|-1}\sigma_{j+1}N_j \leq \frac{\epsilon - (1-\mu/\eta)^{\kappa}\left(F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*)\right)}{\left(1-(1-\mu/\eta)^{\kappa}\right)\frac{\eta^2\Phi}{2\mu D^2}}. \quad (23)$$
2) *$\theta^{(k)}_{L_{j,i}}$, $\forall i,j,k$, given by* (18)*, where the values of $\sigma_1^{(k)}, \cdots, \sigma_{|\mathcal{L}|}^{(k)}$ satisfy* (19) *with δ given by*
$$\delta \geq 1 - \sqrt[\kappa]{\frac{\epsilon}{F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*)}}. \quad (24)$$

Second, we obtain the number of global iterations required to reach a desired accuracy for a predetermined policy of determining the D2D rounds in different clusters (see Appendix E):

Corollary 2. *With ϵ as in Corollary 1, either of the following two conditions give a sufficient number of global iterations κ to achieve $F(\mathbf{w}^{(\kappa)}) - F(\mathbf{w}^*) \leq \epsilon$:*

1) *If the $\theta^{(k)}_{L_{j,i}}$, $\forall i,j,k$, satisfy* (16) *given $\sigma_1, \cdots, \sigma_{|\mathcal{L}|}$, and $\epsilon \geq \frac{\eta^2\Phi}{2\mu D^2}\sum_{j=0}^{|\mathcal{L}|-1}\sigma_{j+1}N_j$, then κ follows* (21)*.*
2) *If the $\theta^{(k)}_{L_{j,i}}$, $\forall i,j,k$, satisfy* (18) *and* (19) *given $\sigma_1^{(k)}, \cdots, \sigma_{|\mathcal{L}|}^{(k)}$ and δ, then κ follows* (22)*.*

4) Varying gradient step size: If we design a time-varying step size β_k that is decreasing over time, we can sharpen the convergence result in Proposition 1, when devices share gradients instead of parameters (see Sec. II-D). In particular, we can guarantee that MH−FL converges to the optimal solution, rather than having a finite optimality gap (see Appendix F):

Proposition 3. *Suppose that the nodes share gradients using the same procedure described in Algorithm 1, and that each parent node samples one of its children uniformly at random. Also, assume that end devices use a step size $\beta_k = \frac{\alpha}{k+\lambda}$, where $\lambda > 1$ and $\alpha > 1/\mu$ at global iteration k, with $\beta_0 \leq 1/\eta$. If the number of D2D rounds are performed according to* (16) *with non-negative constants $\sigma_1, \cdots, \sigma_{|\mathcal{L}|}$, we have*
$$\mathbb{E}[F(\mathbf{w}^{(k)}) - F(\mathbf{w}^*)] \leq \frac{\Gamma}{k+\lambda}, \quad (25)$$
where
$$\Gamma = \max\left\{\lambda(F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*)), \frac{\eta\alpha^2\Phi\sum_{j=0}^{|\mathcal{L}|-1}\sigma_{j+1}N_j}{2D^2(\alpha\mu - 1)}\right\}. \quad (26)$$
Consequently, under such conditions, MH−FL *converges to the optimal solution: $\lim_{k\to\infty}\mathbb{E}[F(\mathbf{w}^{(k)}) - F(\mathbf{w}^*)] = 0$.*

The bound in (25) implies a rate of convergence of $O(1/k)$, which is slower than the linear convergence obtained in Proposition 2, but also allows tapering of the D2D rounds over time as in Proposition 1.

5) Cluster sampling: In a large-scale network with millions of nodes, it may be desirable to reduce upstream communications even further than what is provided by LUT clusters. We develop a *cluster sampling* technique where a portion of the clusters are activated in model training at each global iteration in Appendix G, and extend Theorem 1 to this case. We leave further investigation of this technique to future work.

D. Control Algorithms for Distributed Consensus Tuning

With all else constant, fewer rounds of D2D results in lower power consumption and network load among the devices in LUT clusters. Motivated by this, we develop control algorithms for MH−FL that tune the number of D2D rounds through time (global aggregations) and space (network layers).

1) Adaptive D2D for loss functions satisfying Assumption 1: We are motivated to realize the two D2D consensus round tuning policies that we obtained in Propositions 1 and 2, which we refer to as Policies A and B, respectively. Policy A will provide a finite optimality gap, with tapering of the D2D rounds through time, while Policy B will provide linear convergence to the optimal, with boosting of the D2D rounds through time. We are interested in realizing these two policies in a distributed manner, where the number of D2D rounds for different device clusters are tuned by the corresponding parent nodes in real-time. It is assumed that parent node of C has an estimate on the topology of the cluster, and thus an upper-bound on the spectral radius of its children cluster graph $\lambda_C^{(k)}$, $\forall k$.

According to (16) and (18), for both policies, tuning of the D2D rounds for cluster C requires knowledge of the divergence of parameters $\Upsilon_C^{(k)}$. Also, Policy A requires a set of fixed D2D control parameters σ_j for clusters located in layer L_j, while Policy B requires the global gradient of the broadcast weight $\left\|\nabla F(\mathbf{w}^{(k-1)})\right\|$ and the real-time D2D control parameters $\sigma_j^{(k)}$. In the following, we first derive the divergence of parameters

Algorithm 2: Adaptive D2D round tuning at each cluster

input : Global aggregation count k, tuning parameter $\omega > 1$, cluster index $C = L_{j,i}$.

output : Number of D2D rounds $\theta^{(k)}_{L_{j,i}}$ for the cluster.

Nodes inside the cluster C iteratively compute (a) and (b) of (27).
Parent node of cluster samples one child and computes (27).
if *asymptotic convergence to optimal desired* **then**
    Parent node uses (29) with stored $\|\nabla \widetilde{F(\mathbf{w}^{(k-1)})}\|$ and δ received from the server to compute $\sigma_j^{(k)}$.
    Parent node uses (18) to compute $\theta^{(k)}_{L_{j,i}}$.
else
    Parent node uses the received consensus tuning parameter $\sigma_j^{(k)}$ from the server in (16) to compute $\theta^{(k)}_{L_{j,i}}$.

in a distributed manner. Then, we focus on realizing the other specific parameters for each policy.

Given Definition 1, at cluster C in layer L_j, $1 \leq j \leq |\mathcal{L}|$, the divergence of the parameters can be approximated as[4]

$$\Upsilon^{(k)}_C \approx \max_{q,q' \in \mathcal{C}^{(k)}} \left\{ \|\widetilde{\mathbf{w}}^{(k)}_q\| - \|\widetilde{\mathbf{w}}^{(k)}_{q'}\| \right\} = \underbrace{\max_{q \in \mathcal{C}^{(k)}} \|\widetilde{\mathbf{w}}^{(k)}_q\|}_{(a)} - \underbrace{\min_{q' \in \mathcal{C}^{(k)}} \|\widetilde{\mathbf{w}}^{(k)}_{q'}\|}_{(b)}. \quad (27)$$

To obtain (a) and (b) in a distributed manner, at any given LUT cluster, each node $n \in \mathcal{C}^{(k)}$ first computes the scalar value $\|\widetilde{\mathbf{w}}^{(k)}_n\|$. Nodes then share these scalar values with their neighbors iteratively. In each iteration, each node saves two scalars: the (i) maximum and (ii) minimum values among the received values and the node's current value. It is easy to verify that for any given communication graph $G^{(k)}_C$ among the cluster nodes, once the number of iterations has exceeded the diameter of $G^{(k)}_C$, the saved values at each node will correspond to (a) and (b) for cluster C in (27). The parent node can then sample the value of one of its children to compute (27).

For Policy A, since the values of $\{\sigma_j\}_{j=1}^{|\mathcal{L}|}$ are fixed through time, one option is for the server to tune them once at the beginning of training and distribute them among all the nodes. If satisfaction of a given accuracy ϵ at a certain iteration κ is desired, we use the result of Corollary 1 and obtain the D2D control parameters as the solution of the following max-min optimization problem: $\arg\max_{\{\sigma_j\}_{j=1}^{|\mathcal{L}|}} \min\{N_{j-1}\sigma_j\}$ subject to (23). It can be verified that the solution is given by

$$\sigma^{\star}_j = \frac{\epsilon - (1-\mu/\eta)^{\kappa} \left(F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*)\right)}{\left(1-(1-\mu/\eta)^{\kappa}\right) \frac{\eta^2 \Phi}{2\mu D^2} N_{j-1}|\mathcal{L}|}, \; 1 \leq j \leq |\mathcal{L}|, \quad (28)$$

which can be broadcast at the beginning of training among the nodes. The reason behind the choice of the aforementioned max-min problem is two-fold. First, according to (16), for a given set of divergence of parameters $\Upsilon^{(k)}_C$ across C, fewer numbers of D2D rounds at each layer L_j is associated with larger values of σ_j, so larger values of D2D control parameters are often desired. Second, this choice of objective function results in smaller values of D2D control parameters as we move down the layers (towards the end devices) and the number of nodes increases. This leads to larger D2D rounds and lower errors in the bottom layers, which is desired in practice given the discussion in Sec. III-B that the errors from the bottom layers are propagated and amplified as we move up the layers.

[4]Here, for practical purposes, we use the lower bound of divergence $\big|\, \|\mathbf{a}\| - \|\mathbf{b}\| \,\big| \leq \|\mathbf{a}-\mathbf{b}\|$. The upper bound alternative $\|\mathbf{a}-\mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|$ can be arbitrarily large even when $\mathbf{a} = \mathbf{b}$.

Algorithm 3: Adaptive D2D round tuning at each cluster for non-convex loss functions

input : Tolerable error of aggregations ψ, global aggregation count k, cluster index $C = L_{j,i}$.

output : Number of D2D rounds $\theta^{(k)}_{L_{j,i}}$ for the cluster.

Nodes inside the cluster iteratively compute (a) and (b) of (27).
Parent node of cluster samples one child and computes (27).
Parent node of the cluster computes the required rounds of D2D as follows with $\sigma_j = \psi D^2/(\Phi N_{j-1}|\mathcal{L}|)$:

$$\begin{cases} \theta^{(k)}_{L_{j,i}} \geq \dfrac{\log(\sigma_j) - 2\log\left(\left|\mathcal{L}^{(k)}_{j,i}\right|^{\frac{3}{2}} \Upsilon^{(k)}_{L_{j,i}}\right)}{2\log\left(\lambda^{(k)}_{L_{j,i}}\right)}, & \text{if } \sigma_j \leq |\mathcal{L}^{(k)}_{j,i}|^3 \left(\Upsilon^{(k)}_{L_{j,i}}\right)^2 \\ \theta^{(k)}_{L_{j,i}} \geq 0, & \text{otherwise.} \end{cases} \quad (30)$$

For Policy B, to obtain $\|\nabla \widetilde{F(\mathbf{w}^{(k-1)})}\|$, we use (6) to approximate $\nabla F(\mathbf{w}^{(k-2)})$ as $\nabla F(\mathbf{w}^{(k-2)}) \approx \left(\mathbf{w}^{(k-2)} - \mathbf{w}^{(k-1)}\right)/\beta$. This is an approximation due to the error introduced in the consensus process. Using this, the main server estimates $\|\nabla \widetilde{F(\mathbf{w}^{(k-1)})}\|$ via $\|\nabla \widetilde{F(\mathbf{w}^{(k-1)})}\| = \frac{1}{\omega}\|\nabla F(\mathbf{w}^{(k-2)})\|$, where we introduce tuning parameter $\omega > 1$ based on the intuition that the norm of the gradient should be decreasing over k, and broadcasts it along with $\mathbf{w}^{(k-1)}$. The choice of ω can be viewed as a tradeoff between the number of global aggregations k and the number of D2D rounds $\theta^{(k)}_C$ per aggregation: as ω increases, we tolerate less consensus error in (19), requiring more D2D rounds $\theta^{(k)}_C$ and fewer global iterations k to achieve an accuracy. Then, the cluster heads obtain the $\sigma^{(k)}_j$, $\forall j,k$, according to the following max-min problem: $\arg\max_{\{\sigma^{(k)}_j\}_{j=1}^{|\mathcal{L}|}} \min\{N_{j-1}\sigma^{(k)}_j\}$ subject to (19) for a given δ. It can be verified that the solution is given by

$$\sigma^{(k)\star}_j = \frac{D^2\mu(\mu-\delta\eta)}{\eta^4 \Phi N_{j-1}|\mathcal{L}|} \|\nabla \widetilde{F(\mathbf{w}^{(k-1)})}\|^2, 1 \leq j \leq |\mathcal{L}|, \forall k. \quad (29)$$

The parameter δ can be tuned by the main server at the beginning of training to guarantee a desired linear convergence, or it can be tuned by (24) to satisfy a desired accuracy at a certain global iteration. In both cases, the server broadcasts this parameter among the nodes at the beginning of training. With δ and $\|\nabla \widetilde{F(\mathbf{w}^{(k-1)})}\|$ in hand, along with the ML model characteristics (D, μ, η) and networked related parameters $(\Phi, |\mathcal{L}|, N_{j-1})$ that can be once broadcast by the server, all the parent nodes of the clusters can calculate (29) at each global iteration, which then can be used in (18) to tune the number of D2D rounds for the children nodes in real-time.

A summary of this procedure for tuning the D2D rounds at a cluster is given in Algorithm 2. In the full `MH-FL` method described in Algorithm 1, this is (optionally) called for each cluster in the lines marked via $**$.

2) Adaptive D2D tuning for non-convex loss functions: Some contemporary ML models, such as neural networks, possess non-convex loss functions for which Assumption 1 does not apply. In these cases, we can develop a heuristic approach to tune the D2D rounds of `MH-FL` if we specify

a maximum tolerable error of aggregations ψ at each global iteration. The resulting procedure is given in Algorithm 3, which is called once for each cluster in Algorithm 1 in place of Algorithm 2 (see the lines marked ##). To execute this, prior to the start of training, each parent node should receive ψ and the number of nodes in its layer. Using Algorithm 3, we can show that the 2-norm of aggregation errors is always bounded by parameter ψ (see Appendix H for the proof).

IV. Experimental Evaluation

We conducted extensive numerical experiments to evaluate MH−FL. In this section, we present the setup and results for a popular dataset and a sample fog topology. Additional results on more datasets and topologies can be found in Appendix I.

A. Experimental Setup

We consider a fog network consisting of a main server and three subsequent layers. There are 125 edge devices in the bottom layer (L_3), clustered into groups of 5 nodes. Each of these clusters communicates with one of 25 parent nodes in layer L_2. The nodes at this layer are in turn clustered into groups of 5, with 5 parent nodes at layer L_1 that communicate with the main server. We consider the cases where (i) all clusters are configured to operate in LUT mode and (ii) all are EUT, which allows us to evaluate the performance differences in terms of model convergence, energy consumption, and parameters transferred. In the LUT case, network topology within clusters follows a random geometric graph [51]. See Appendix I for the detailed discussion of our implementation.

We consider a 10-class image classification task on the standard MNIST dataset of 70K handwritten digits (http://yann.lecun.com/exdb/mnist/). We evaluate with both regularized SVM and fully-connected neural network (NN) classifiers as loss functions; SVM satisfies Assumption 1 while NN does not. Unless stated otherwise, the results are presented using SVM. Samples are distributed across devices in either an i.i.d. or non-i.i.d. manner; for i.i.d., each device has samples from each class, while for non-i.i.d., each device has samples from only of the 10 classes. More details on the dataset, classifiers, and hyperparameter tuning procedure are available in Appendix I: there, we also provide additional results on the Fashion MNIST (F-MNIST) dataset and for a network of 625 edge devices.

B. MH−FL with Fixed Consensus Rounds

1) MH−FL *with fixed step size:* We consider a scenario in which the number of D2D rounds is set to be a constant value θ over all clusters, which provides useful insights for the rest of the results. In Fig. 5, we compare the performance of MH−FL when all the clusters work in LUT mode and have fixed rounds of D2D with the case where the network consists of all EUT clusters (referred to as "EUT baseline"). The EUT case is identical (in terms of convergence) to carrying out centralized gradient descent over the entire dataset. We see that increasing the number of consensus at different layers increases the accuracy and stability of the model training for MH−FL. Although convergence is not achieved in all cases (in particular, when θ is 1 and 2), if the number of D2D rounds chosen is larger than 15, comparable performance to EUT is achieved. This performance is characterized by linear convergence, as can be seen in Fig. 5(c) with logarithmic axis scales.

2) MH−FL *with decaying step size:* The effect of decreasing the gradient descent step size (Proposition 3) is depicted in Fig. 6. This verifies that the decay can suppress the finite optimality bound and provide convergence to the optimal model. Also, the convergence occurs at a slower pace compared with the baseline, which is in line with our theoretical results (convergence rate of $O(1/k)$). Fig. 6 further reveals the inherent trade-off between conducting a higher number of D2D rounds with a constant learning rate (higher power consumption from more rounds, but with a fast convergence speed) and performing a fewer number of D2D rounds with a decaying learning rate (lower power consumption with a slower convergence).

C. MH−FL with Adaptive D2D Round Tuning

We next study the case when our distributed D2D tuning scheme is utilized. The results are depicted for both convergence cases, i.e., where a finite optimality gap is tolerable (Figs. 7, 8) and when the linear convergence to the optimal solution is desired (Figs. 9, 10). Recall that Propositions 1&2 obtain the sufficient number of D2D rounds based on an upper bound; for this experiment, we observed that $\log(\sigma_j)$ and $\log(\sigma_j^{(k)})$ in (16) and (18) can be scaled and used as $\log(\chi\sigma_j^{(k)})$ and $\log(\chi\sigma_j^{(k)})$, $\forall j$, where $\chi \in [1, 15]$ to obtain fewer rounds of D2D while satisfying the desired convergence behavior.

1) MH−FL *with finite optimality gap:* Fig. 7 depicts the result for the case where local datasets are i.i.d., with the values of $\{\sigma_j\}_{j=1}^{|\mathcal{L}|}$ depicted. In the figures, $\overline{\theta^{(k)}}$ denotes the average number of D2D rounds employed by clusters over all the network layers at global iteration k, and $\overline{\theta_{L_j}^{(k)}}$ denotes the average number of D2D rounds at iteration k in layer L_j. We observe that (i) the D2D rounds performed in the network is tapered through time (subplot c), and (ii) the D2D rounds performed at different network layers is tapered through space, where higher layers of the network perform fewer rounds (subplots d-f). We perform the same experiment with non-i.i.d datasets across the nodes in Fig. 8. Comparing Fig. 8 to 7, it can be observed that non-i.i.d. introduces oscillations on the number of D2D performed at different network layers, and the smaller values of D2D control parameters result in larger numbers of D2D rounds which leads to more stable training.

2) MH−FL *with linear convergence:* The same experiment is repeated in Figs. 9, 10 for the linear convergence case. We see that the number of D2D rounds is tapered through space (subplots d-f) and is boosted over time index k (subplots c-f). This is due to the decrease in the norm of gradient in the right hand side of (29) over time, which calls for an increment in the number of D2D rounds. Comparing Figs. 9 and 10 with Figs. 7 and 8, we see that guaranteeing the linear convergence comes with the tradeoff of performing a larger number of D2D rounds over time. In Figs. 7, 8, a small optimality gap is achieved, while the number of D2D rounds is tapered over time.

3) MH−FL *with adaptive rounds of D2D for non-convex ML models:* Recall that we developed Algorithm 3 for non-convex ML loss functions. Figs. 11 and 12 give results with NNs for different values of tolerable error of aggregations ψ, under i.i.d and non-i.i.d data distributions, respectively. These figures show the effect of tolerable error of aggregations on the performance of NNs; by decreasing the tolerable error, the

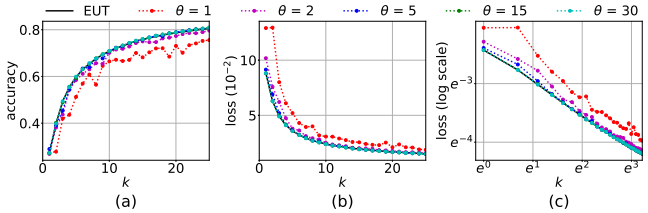


Fig. 5: Performance comparison between baseline EUT and MH−FL when a fixed number of D2D rounds θ is used at every cluster in the network, for non-i.i.d data. As the number of D2D rounds increases, MH−FL performs more similar to the EUT baseline and the learning becomes more stable.

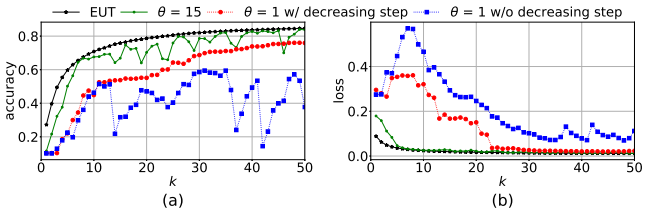


Fig. 6: Performance comparison between baseline EUT, and MH−FL with and without (w/o) decreasing the gradient descent step size. Decreasing the step size can provide convergence to the optimal solution in cases where a fixed step size is not capable, but also has a slower convergence speed.

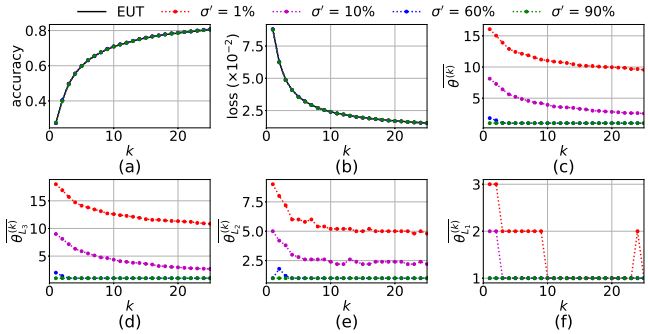


Fig. 7: Performance comparison between baseline EUT and MH−FL for i.i.d data when a finite optimality gap is tolerable. σ_j at L_j is fixed as $\sigma_j = \sigma' \max_i \Upsilon^{(1)}_{L_{j,i}}$. The tapering of D2D rounds through time (c) and space (layers) (d)-(f) can be observed.

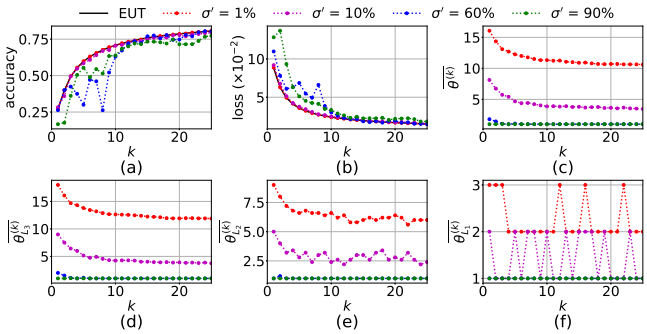


Fig. 8: Performance comparison between baseline EUT and MH−FL for non-i.i.d data when a finite optimality gap is tolerable. σ_i is set as in Fig. 7. Smaller loss and higher accuracy are achieved with smaller σ', implying more rounds of D2D are required.

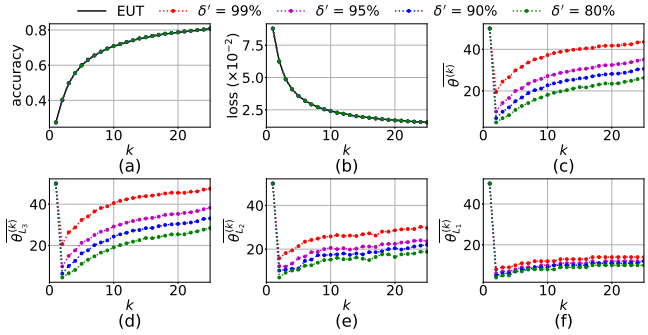


Fig. 9: Performance comparison between baseline EUT and MH−FL for i.i.d data when linear convergence to the optimal is desired. The value of δ is set at $\delta = \delta' \frac{\mu}{\eta}$. Boosting of the D2D rounds through time can be observed in (c)-(f) as k is increased. Also, tapering through space can be observed by comparing the D2D rounds across the bottom subplots.

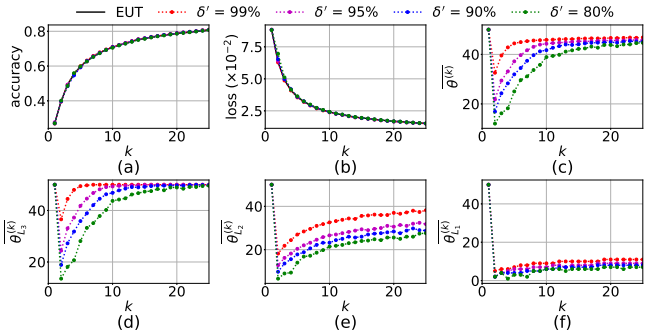


Fig. 10: Performance comparison between baseline EUT and MH−FL for non-i.i.d data when linear convergence to the optimal is desired. The value of δ is set as in Fig. 9. Smaller values of loss and higher accuracy are both associated with larger value of δ, which results in lower error tolerance and more rounds of D2D.

number of D2D rounds is increased, and the performance is enhanced. These figures also reveal a tapering of the number of consensus through time and space in the i.i.d case.

4) Analytical vs. experimental bound comparison: We investigate the number of aggregations required to obtain a certain accuracy under linear convergence (Corollary 1). In Fig. 13, we compare the result obtained from Policy B using (29) to that observed in our experiments. The results indicate that the theoretical bounds are reasonably tight.

D. Network Resource Utilization

We now study the network resource utilization of MH−FL. In particular, we consider two metrics: (i) the amount of data transferred between the network layers, and (ii) the accumulated energy consumption of the edge devices. In both cases, EUT and MH−FL are trained to reach 98% of the final accuracy achieved after 50 iterations of centralized gradient descent. We consider four scenarios, corresponding to those used in Figs. 7, 8, 11, 12. To obtain the accumulated energy, we consider the transmission power of end devices to be 10dbm in D2D and 24dbm in uplink mode [52], [53], and assume that transmission of parameters at each round occurs with data rate of 1Mbits/s with 32-bit quantization per model parameter element. The accumulated energy consumption of the edge devices through the training phase is depicted in Fig. 14, which reveals around 50% energy saving on average as compared to the EUT baseline. The accumulated number of parameters transferred over the network layers are shown in Fig. 15, revealing 80% reduction in the number of parameters transferred over the layers as compared to the baseline. We conduct further numerical studies in Appendix I-C to reveal the impact of our D2D control parameters $\{\sigma_j\}_{j=1}^{|\mathcal{L}|}$ and our tolerable aggregation error ψ on the energy and transmit parameters savings.

V. Conclusion and Future Work

We developed multi-stage hybrid federated learning (MH−FL), a novel methodology which migrates the star topology of

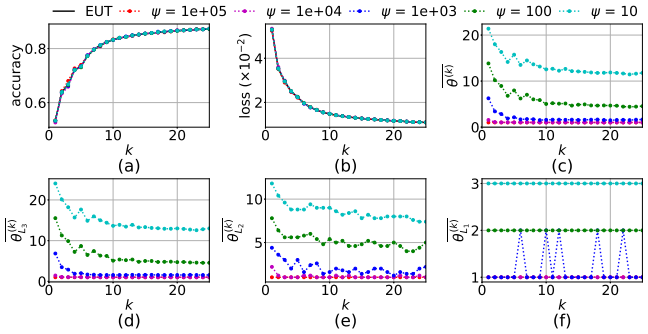


Fig. 11: Performance comparison between baseline EUT and MH−FL under i.i.d data using NNs with different values of ψ. Tapering the D2D rounds through time can be observed. Also, tapering through space can be observed by comparing the D2D rounds across the bottom subplots.

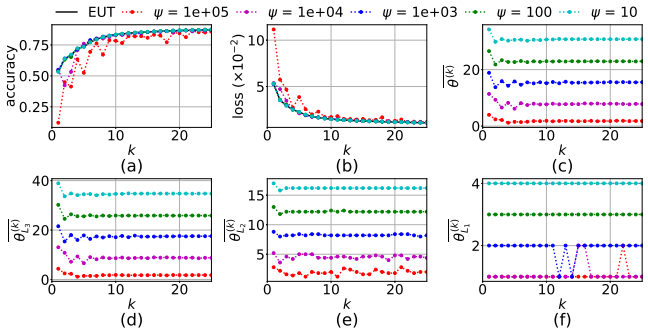


Fig. 12: Performance comparison between baseline EUT and MH−FL under non-i.i.d data using NNs with different values of ψ. Lower loss and higher accuracy are associated with smaller values of ψ, which result in lower error tolerance and larger numbers of D2D rounds.

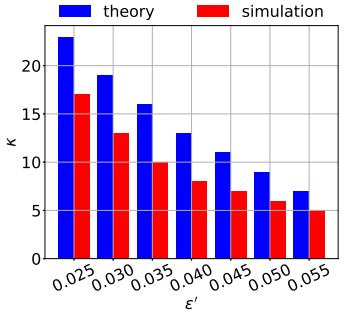


Fig. 13: Theoretical vs. simulation results regarding the number of global iterations to achieve an accuracy of $\epsilon'(F(\mathbf{w}^{(0)})-F(\mathbf{w}^*))$ for different ϵ'. Convergence in practice is faster than the derived upper bound.

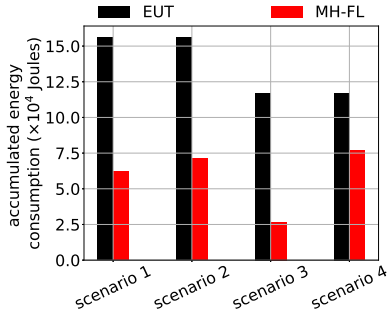


Fig. 14: Accumulated energy consumption of EUT and MH−FL over scenario 1: $\sigma' = 0.1$ from Fig. 7, scenario 2: $\sigma' = 0.1$ from Fig. 8, scenario 3: $\psi = 10^4$ from Fig. 11, and scenario 4: $\psi = 10^4$ from Fig. 12.

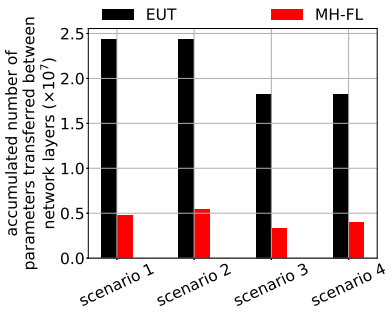


Fig. 15: Comparison of parameters transferred among layers in EUT vs MH−FL over scenario 1: $\sigma' = 0.1$ from Fig. 7, scenario 2: $\sigma' = 0.1$ from Fig. 8, scenario 3: $\psi = 10^4$ from Fig. 11, and scenario 4: $\psi = 10^4$ from Fig. 12.

federated learning to a multi-layer cluster-based distributed architecture incorporating cooperative D2D communications, which constitutes a semi-decentralized learning architecture. We theoretically obtained the convergence bound of MH−MT explicitly considering the time varying network topology, time varying number of D2D rounds at different network clusters, and inherent ML model characteristics. We proposed a set of policies for the number of D2D rounds conducted at different network clusters under which convergence either to a finite optimality gap or the global optimum can be achieved. We further used these policies to develop a set of adaptive distributed control algorithms that tune the number of D2D rounds at different clusters of the network in real-time.

This paper motivates several directions for future work. Investigating more specific system characteristics that have been considered for federated learning – such as communication imperfections, interference management, mitigation of stragglers, and device scheduling – for the multi-stage hybrid structure of fog networks is promising. Also, the proposed network dimension of federated learning motivates works on network (re-)formation, congestion-aware data transfer and load balancing, and topology design for performance optimization. Furthermore, integrating the recently proposed asynchronous federated learning paradigm [54] with the semi-decentralized architecture proposed in this paper is an interesting direction. Finally, in this work, we have assumed that the operation of device clusters as LUT vs. EUT is provided as an input to our methodology; namely, by the physical/link-layer protocols in place, where D2D communication links are established. A holistic, cross-layer optimization approach that jointly optimizes model training and resource utilization metrics over the partitioning of devices into LUT vs. EUT and the subsequent operation of LUT clusters is another promising future direction.

REFERENCES

[1] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proc. IEEE Conf. Comput. Vision Pattern Recog. (CVPR)*, 2012, pp. 3642–3649.

[2] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. Int. Conf. Mach. Learn.*, 2008, pp. 160–167.

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Stat.*, 2017, pp. 1273–1282.

[4] S. Hosseinalipour, C. G. Brinton, V. Aggarwal, H. Dai, and M. Chiang, "From federated to fog learning: Distributed machine learning over heterogeneous wireless networks," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 41–47, 2020.

[5] M. Chiang, S. Ha, F. Risso, T. Zhang, and I. Chih-Lin, "Clarifying fog computing and networking: 10 questions and answers," *IEEE Commun. Mag.*, vol. 55, no. 4, pp. 18–20, 2017.

[6] M. N. Tehrani, M. Uysal, and H. Yanikomeroglu, "Device-to-device communication in 5G cellular networks: challenges, solutions, and future directions," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 86–92, 2014.

[7] M. Abolhasan, T. Wysocki, and E. Dutkiewicz, "A review of routing protocols for mobile ad hoc networks," *Ad hoc Netw.*, vol. 2, no. 1, pp. 1–22, 2004.

[8] S. Zeadally, R. Hunt, Y.-S. Chen, A. Irwin, and A. Hassan, "Vehicular ad hoc networks (VANETS): status, results, and challenges," *Telecommun. Syst.*, vol. 50, no. 4, pp. 217–241, 2012.

[9] I. Bekmezci, O. K. Sahingoz, and Ş. Temel, "Flying ad-hoc networks (FANETs): A survey," *Ad hoc Netw.*, vol. 11, no. 3, pp. 1254–1270, 2013.

[10] K. Akkaya and M. Younis, "A survey on routing protocols for wireless sensor networks," *Ad hoc Netw.*, vol. 3, no. 3, pp. 325–349, 2005.

[11] Y. Zhang, D. J. Love, J. V. Krogmeier, C. R. Anderson, R. W. Heath, and D. R. Buckmaster, "Challenges and opportunities of future rural wireless communications," *IEEE Commun. Mag.*, 2021.

[12] S. Maheshwari, D. Raychaudhuri, I. Seskar, and F. Bronzino, "Scalability and performance evaluation of edge cloud systems for latency constrained applications," in *IEEE/ACM Symp. Edge Comput.*, 2018, pp. 286–299.

[13] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *arXiv preprint arXiv:1909.07972*, 2019.

[14] N. H. Tran, W. Bao, A. Zomaya, M. N. H. Nguyen, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, 2019, pp. 1387–1395.

[15] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3546–3557, 2020.

[16] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "Federated learning with quantization constraints," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, 2020, pp. 8851–8855.

[17] C.-S. Lee, N. Michelusi, and G. Scutari, "Finite-bit quantization for distributed algorithms with linear convergence," *arXiv preprint arXiv:2107.11304*, 2021.

[18] C. Renggli, S. Ashkboos, M. Aghagolzadeh, D. Alistarh, and T. Hoefler, "SparCML: High-performance sparse communication for machine learning," in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.*, 2019, pp. 1–15.

[19] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun. (JSAC)*, vol. 37, no. 6, pp. 1205–1221, 2019.

[20] D. Ye, R. Yu, M. Pan, and Z. Han, "Federated learning in vehicular edge computing: A selective model aggregation approach," *IEEE Access*, vol. 8, pp. 23 920–23 935, 2020.

[21] Y. Tu, Y. Ruan, S. Wang, S. Wagle, C. G. Brinton, and C. Joe-Wang, "Network-aware optimization of distributed learning for fog computing," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, 2020.

[22] S. Wang, M. Lee, S. Hosseinalipour, R. Morabito, M. Chiang, and C. G. Brinton, "Device sampling for heterogeneous federated learning: Theory, algorithms, and implementation," in *IEEE Conf. Comput. Commun. (INFOCOM)*, 2021, pp. 1–10.

[23] S. Ji, W. Jiang, A. Walid, and X. Li, "Dynamic sampling and selective masking for communication-efficient federated learning," *arXiv preprint arXiv:2003.09603*, 2020.

[24] W. Wang, Y. Sun, B. Eriksson, W. Wang, and V. Aggarwal, "Wide compression: Tensor ring nets," in *Proc. IEEE Conf. Comput. Vision Pattern Recog. (CVPR)*, 2018, pp. 9329–9338.

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

[26] 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.*, pp. 1–1, 2020.

[27] M. S. H. Abad, E. Ozfatura, D. GUndUz, and O. Ercetin, "Hierarchical federated learning across heterogeneous cellular networks," in *Proc. IEEE ICASSP*, 2020, pp. 8866–8870.

[28] S. Savazzi, M. Nicoli, and V. Rampa, "Federated learning with cooperating devices: A consensus approach for massive IoT networks," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 4641–4654, 2020.

[29] C. Hu, J. Jiang, and Z. Wang, "Decentralized federated learning: A segmented gossip approach," *arXiv preprint arXiv:1908.07782*, 2019.

[30] A. Elgabli, J. Park, A. S. Bedi, M. Bennis, and V. Aggarwal, "GADMM: Fast and communication efficient framework for distributed machine learning," *arXiv preprint arXiv:1909.00047*, 2019.

[31] V. Smith, S. Forte, C. Ma, M. Takáč, M. I. Jordan, and M. Jaggi, "CoCoA: A general framework for communication-efficient distributed optimization," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 8590–8638, 2017.

[32] P. Richtárik and M. Takáč, "Distributed coordinate descent method for learning with big data," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2657–2681, 2016.

[33] S. Niknam, H. S. Dhillon, and J. H. Reed, "Federated learning for wireless communications: Motivation, opportunities and challenges," *arXiv preprint arXiv:1908.06847*, 2019.

[34] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Toward an intelligent edge: Wireless communication meets machine learning," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 19–25, 2020.

[35] C. L. P. Chen, G. Wen, Y. Liu, and F. Wang, "Adaptive consensus control for a class of nonlinear multiagent time-delay systems using neural networks," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 25, no. 6, pp. 1217–1226, 2014.

[36] T. Li and J.-F. Zhang, "Consensus conditions of multi-agent systems with time-varying topologies and stochastic communication noises," *IEEE Trans. Autom. Control*, vol. 55, no. 9, pp. 2043–2057, 2010.

[37] S. Kar and J. M. F. Moura, "Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 355–369, 2009.

[38] S. Manfredi, "Design of a multi-hop dynamic consensus algorithm over wireless sensor networks," *Control Eng. Practice*, vol. 21, no. 4, pp. 381–394, 2013.

[39] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Auto. Control*, vol. 54, no. 1, pp. 48–61, 2009.

[40] T. Chang, M. Hong, and X. Wang, "Multi-agent distributed optimization via inexact consensus ADMM," *IEEE Trans. Signal Process.*, vol. 63, no. 2, pp. 482–497, 2015.

[41] T. Chang, A. Nedić, and A. Scaglione, "Distributed constrained optimization by consensus-based primal-dual perturbation method," *IEEE Trans. Autom. Control*, vol. 59, no. 6, pp. 1524–1538, 2014.

[42] B. Johansson, T. Keviczky, M. Johansson, and K. H. Johansson, "Subgradient methods and consensus algorithms for solving convex optimization problems," in *Proc. IEEE Conf. Decis. Control*, 2008, pp. 4185–4190.

[43] R. N. Clarke, "Expanding mobile wireless capacity: The challenges presented by technology and economics," *Telecommun. Policy*, vol. 38, no. 8-9, pp. 693–708, 2014.

[44] U. N. Kar and D. K. Sanyal, "An overview of device-to-device communication in cellular networks," *ICT Express*, vol. 4, no. 4, pp. 203–208, 2018.

[45] T. Zeng, O. Semiari, M. Mozaffari, M. Chen, W. Saad, and M. Bennis, "Federated learning in the sky: Joint power allocation and scheduling with UAV swarms," *arXiv preprint arXiv:2002.08196*, 2020.

[46] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Syst. & Control Lett.*, vol. 53, no. 1, pp. 65–78, 2004.

[47] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization," *arXiv:1909.13014*, 2019.

[48] C. Dinh, N. H. Tran, M. N. Nguyen, C. S. Hong, W. Bao, A. Zomaya, and V. Gramoli, "Federated learning over wireless networks: Convergence analysis and resource allocation," *arXiv preprint arXiv:1910.13067*, 2019.

[49] M. P. Friedlander and M. Schmidt, "Hybrid deterministic-stochastic methods for data fitting," *SIAM J. Sci. Comput.*, vol. 34, no. 3, pp. A1380–A1405, 2012.

[50] L. Xiao, S. Boyd, and S.-J. Kim, "Distributed average consensus with least-mean-square deviation," *J. Parallel Distrib. Comput.*, vol. 67, no. 1, pp. 33–46, 2007.

[51] X. Jia, "Wireless networks and random geometric graphs," in *Proc. Int. Symp. Parallel Arch. Alg. Netw.*, 2004, pp. 575–579.

[52] M. Hmila, M. Fernández-Veiga, M. Rodríguez-Pérez, and S. Herrería-Alonso, "Energy efficient power and channel allocation in underlay device to multi device communications," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5817–5832, 2019.

[53] S. Dominic and L. Jacob, "Joint resource block and power allocation through distributed learning for energy efficient underlay D2D communication with rate guarantee," *Comput. Commun.*, 2020.

[54] C. Xie, S. Koyejo, and I. Gupta, "Asynchronous federated optimization," *arXiv preprint arXiv:1903.03934*, 2019.

[55] B. T. Polyak, "Gradient methods for minimizing functionals," *Zh. Vychisl. Mat. Mat. Fiz*, vol. 3, no. 4, pp. 643–653, 1963.

[56] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic programming*. Athena Scientific, 1996.

Seyyedali Hosseinalipour (M'20) received his Ph.D. in EE from NCSU in 2020. He is currently a postdoctoral researcher at Purdue University.

Sheikh Shams Azam is a Ph.D. student at Purdue University. He received his B.Tech. in ECE from NITK, India in 2015.

Christopher G. Brinton (SM'20) is an Assistant Professor of ECE at Purdue University. He received his Ph.D. in EE from Princeton University in 2016.

Nicolò Michelusi (SM'19) received his Ph.D. in EE from University of Padova, Italy, in 2013. He is an Assistant Professor at Arizona State University.

Vaneet Aggarwal (SM'15) received his Ph.D. in EE from Princeton University in 2010. He is currently an Associate Professor at Purdue University.

David Love (F'15) is the Nick Trbovich Professor of ECE at Purdue University. He received the Ph.D. degree in EE from University of Texas at Austin in 2004.

Huaiyu Dai (F'17) received the Ph.D. degree in EE from Princeton University in 2002. He is currently a Professor of ECE at NCSU, holding the title of University Faculty Scholar.

APPENDIX A
PROOF OF THEOREM 1

Proof. We carry out the proof in three parts: In part I, we obtain the convergence behavior of MH−FL given arbitrary aggregation errors at the sampled devices. In part II, we obtain the aggregation error caused by the D2D consensus process. Finally, in part III, we derive the final convergence bound.

A. PART I: Convergence Bound for General Local Aggregation Error at the Sampled Nodes

We first aim to bound the per-iteration decrease of the gap between the function $F(\mathbf{w}^{(k)})$ and $F(\mathbf{w}^*)$. Using the Taylor expansion and the η-smoothness of function F, the following quadratic upper-bound can be obtained:

$$F(\mathbf{w}^{(k)}) \leq F(\mathbf{w}^{(k-1)}) + (\mathbf{w}^{(k)} - \mathbf{w}^{(k-1)})^\top \nabla F(\mathbf{w}^{(k-1)}) + \frac{\eta}{2}\left\|\mathbf{w}^{(k)} - \mathbf{w}^{(k-1)}\right\|^2, \quad \forall k. \tag{31}$$

To find the relationship between $\mathbf{w}^{(k-1)}$ and $\mathbf{w}^{(k)}$, we follow the procedure described in the main text. For parent node a_p, let a'_{p+1} denote the corresponding sampled node, $\forall p$, e.g., in the following nested sums $a'_{|\mathcal{L}|}$ denotes the sampled node in the last layer by parent node $a_{|\mathcal{L}|-1}$ in its above layer. This corresponds to an arbitrary realization of the children sampling at different parent nodes. Let a'_1 denote the sampled node by the main server in L_1. The model parameter of this node is given by

$$\begin{aligned}
\widehat{\mathbf{w}}^{(k)}_{a'_1} = &\frac{\sum_{a_1\in\mathcal{L}^{(k)}_{1,1}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}|\mathcal{D}_{a_{|\mathcal{L}|}}|\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}}}{|\mathcal{L}^{(k)}_{1,1}|}\\
&-\frac{\sum_{a_1\in\mathcal{L}^{(k)}_{1,1}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\beta|\mathcal{D}_{a_{|\mathcal{L}|}}|\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})}{|\mathcal{L}^{(k)}_{1,1}|}\\
&+\sum_{a_1\in\mathcal{L}^{(k)}_{1,1}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})}\frac{\mathbb{1}^{(k)}_{\{Q(a_{|\mathcal{L}|-1})\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})|\mathbf{c}^{(k)}_{a'_{|\mathcal{L}|}}}{|\mathcal{L}^{(k)}_{1,1}|}\\
&+\sum_{a_1\in\mathcal{L}^{(k)}_{1,1}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})}\frac{\mathbb{1}^{(k)}_{\{Q(a_{|\mathcal{L}|-2})\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})|\mathbf{c}^{(k)}_{a'_{|\mathcal{L}|-1}}}{|\mathcal{L}^{(k)}_{1,1}|}+\\
&\vdots\\
&+\sum_{a_1\in\mathcal{L}^{(k)}_{1,1}}\frac{\mathbb{1}^{(k)}_{\{Q(a_1)\}}|\mathcal{Q}^{(k)}(a_1)|\mathbf{c}^{(k)}_{a'_2}}{|\mathcal{L}^{(k)}_{1,1}|}+\mathbb{1}^{(k)}_{\{L_{1,1}\}}\mathbf{c}^{(k)}_{a'_1},
\end{aligned} \tag{32}$$

which the main server uses to obtain the next global parameter as follows (due to the existence of the indicator function in the last term of the above expression, the following expression holds regardless of the operating mode of the cluster at layer L_1):

$$\mathbf{w}^{(k)} = \frac{|\mathcal{L}^{(k)}_{1,1}|\widehat{\mathbf{w}}^{(k)}_{a'_1}}{D}. \tag{33}$$

Based on (3), it can be verified that

$$\sum_{a_1\in\mathcal{L}^{(k)}_{1,1}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}|\mathcal{D}_{a_{|\mathcal{L}|}}|\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}} = D\mathbf{w}^{(k-1)}. \tag{34}$$

Also, using (1), we have

$$\sum_{a_1\in\mathcal{L}^{(k)}_{1,1}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\beta|\mathcal{D}_{a_{|\mathcal{L}|}}|\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}}) = \beta D\nabla F(\mathbf{w}^{(k-1)}). \tag{35}$$

Replacing the above two equations in (32) and performing the update given by (33), we get

$$\mathbf{w}^{(k)} = \mathbf{w}^{(k-1)} - \beta\nabla F(\mathbf{w}^{(k-1)}) + \frac{1}{D}\Bigg(\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \sum_{a_2\in\mathcal{Q}^{(k)}(a_1)} \sum_{a_3\in\mathcal{Q}^{(k)}(a_2)} \cdots \sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})} \mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-1})\right\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})| \mathbf{c}^{(k)}_{a'_{|\mathcal{L}|}}$$
$$+ \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \sum_{a_2\in\mathcal{Q}^{(k)}(a_1)} \sum_{a_3\in\mathcal{Q}^{(k)}(a_2)} \cdots \sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})} \mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-2})\right\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})| \mathbf{c}^{(k)}_{a'_{|\mathcal{L}|-1}} + \tag{36}$$
$$\vdots$$
$$+ \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \mathbb{1}^{(k)}_{\{Q(a_1)\}} |\mathcal{Q}^{(k)}(a_1)| \mathbf{c}^{(k)}_{a'_2} + \mathbb{1}^{(k)}_{\{L_{1,1}\}} |\mathcal{L}^{(k)}_{1,1}| \mathbf{c}^{(k)}_{a'_1} \Bigg),$$

Calculating $\mathbf{w}^{(k)} - \mathbf{w}^{(k-1)}$ using the above equation and replacing the result in (31) yields

$$F(\mathbf{w}^{(k)}) - F(\mathbf{w}^{(k-1)}) \leq \left(\frac{\eta\beta^2}{2} - \beta\right) \left\|\nabla F(\mathbf{w}^{(k-1)})\right\|^2$$
$$+ \left(\frac{1-\beta\eta}{D}\right) \Bigg(\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \sum_{a_2\in\mathcal{Q}^{(k)}(a_1)} \sum_{a_3\in\mathcal{Q}^{(k)}(a_2)} \cdots \sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})} \mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-1})\right\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})| \mathbf{c}^{(k)}_{a'_{|\mathcal{L}|}}$$
$$+ \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \sum_{a_2\in\mathcal{Q}^{(k)}(a_1)} \sum_{a_3\in\mathcal{Q}^{(k)}(a_2)} \cdots \sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})} \mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-2})\right\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})| \mathbf{c}^{(k)}_{a'_{|\mathcal{L}|-1}} + \cdots$$
$$+ \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \mathbb{1}^{(k)}_{\{Q(a_1)\}} |\mathcal{Q}^{(k)}(a_1)| \mathbf{c}^{(k)}_{a'_2} + \mathbb{1}^{(k)}_{\{L_{1,1}\}} |\mathcal{L}^{(k)}_{1,1}| \mathbf{c}^{(k)}_{a'_1} \Bigg)^{\top} \nabla F(\mathbf{w}^{(k-1)}) \tag{37}$$
$$+ \frac{\eta}{2D^2} \Bigg\| \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \sum_{a_2\in\mathcal{Q}^{(k)}(a_1)} \sum_{a_3\in\mathcal{Q}^{(k)}(a_2)} \cdots \sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})} \mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-1})\right\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})| \mathbf{c}^{(k)}_{a'_{|\mathcal{L}|}}$$
$$+ \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \sum_{a_2\in\mathcal{Q}^{(k)}(a_1)} \sum_{a_3\in\mathcal{Q}^{(k)}(a_2)} \cdots \sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})} \mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-2})\right\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})| \mathbf{c}^{(k)}_{a'_{|\mathcal{L}|-1}} + \cdots$$
$$+ \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \mathbb{1}^{(k)}_{\{Q(a_1)\}} |\mathcal{Q}^{(k)}(a_1)| \mathbf{c}^{(k)}_{a'_2} + \mathbb{1}^{(k)}_{\{L_{1,1}\}} |\mathcal{L}^{(k)}_{1,1}| \mathbf{c}^{(k)}_{a'_1} \Bigg\|^2.$$

Tuning the learning rate as $\beta = \frac{1}{\eta}$, we obtain

$$F(\mathbf{w}^{(k)}) - F(\mathbf{w}^{(k-1)}) \leq \frac{-1}{2\eta} \left\|\nabla F(\mathbf{w}^{(k-1)})\right\|^2 +$$
$$\frac{\eta}{2D^2} \Bigg\| \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \sum_{a_2\in\mathcal{Q}^{(k)}(a_1)} \sum_{a_3\in\mathcal{Q}^{(k)}(a_2)} \cdots \sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})} \mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-1})\right\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})| \mathbf{c}^{(k)}_{a'_{|\mathcal{L}|}}$$
$$+ \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \sum_{a_2\in\mathcal{Q}^{(k)}(a_1)} \sum_{a_3\in\mathcal{Q}^{(k)}(a_2)} \cdots \sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})} \mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-2})\right\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})| \mathbf{c}^{(k)}_{a'_{|\mathcal{L}|-1}} + \tag{38}$$
$$\vdots$$
$$+ \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \mathbb{1}^{(k)}_{\{Q(a_1)\}} |\mathcal{Q}^{(k)}(a_1)| \mathbf{c}^{(k)}_{a'_2} + \mathbb{1}^{(k)}_{\{L_{1,1}\}} |\mathcal{L}^{(k)}_{1,1}| \mathbf{c}^{(k)}_{a'_1} \Bigg\|^2.$$

Considering the right hand side of the above inequality, to bound $\left\|\nabla F(\mathbf{w}^{(k-1)})\right\|^2$, we use the strong convexity property of F. Considering the strong convexity criterion in Assumption 1 with x replaced by $\mathbf{w}^{(k-1)}$ and minimizing the both hand sides, the minimum of the left hand side occurs at $y = \mathbf{w}^*$ and the minimum of the right hand side occurs at $y = \mathbf{w}^{(k-1)} - \frac{1}{\mu}\nabla F(\mathbf{w}^{(k-1)})$. Replacing these values in the strong convexity criterion in Assumption 1 results in Polyak-Lojasiewicz inequality [55] in the following form:

$$\left\|\nabla F(\mathbf{w}^{(k-1)})\right\|^2 \geq (F(\mathbf{w}^{(k-1)}) - F(\mathbf{w}^*))2\mu, \tag{39}$$

which yields

$$
\begin{aligned}
&F(\mathbf{w}^{(k)}) - F(\mathbf{w}^{(k-1)}) \leq \frac{-\mu}{\eta}(F(\mathbf{w}^{(k-1)}) - F(\mathbf{w}^*)) + \\
&\frac{\eta}{2D^2}\Bigg[\Bigg\| \sum_{a_1 \in \mathcal{L}_{1,1}^{(k)}} \sum_{a_2 \in \mathcal{Q}^{(k)}(a_1)} \sum_{a_3 \in \mathcal{Q}^{(k)}(a_2)} \cdots \sum_{a_{|\mathcal{L}|-1} \in \mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})} \mathbb{1}^{(k)}_{\{Q(a_{|\mathcal{L}|-1})\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})| \mathbf{c}^{(k)}_{a'_{|\mathcal{L}|}} \\
&+ \sum_{a_1 \in \mathcal{L}_{1,1}^{(k)}} \sum_{a_2 \in \mathcal{Q}^{(k)}(a_1)} \sum_{a_3 \in \mathcal{Q}^{(k)}(a_2)} \cdots \sum_{a_{|\mathcal{L}|-2} \in \mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})} \mathbb{1}^{(k)}_{\{Q(a_{|\mathcal{L}|-2})\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})| \mathbf{c}^{(k)}_{a'_{|\mathcal{L}|-1}} + \\
&\vdots \\
&+ \sum_{a_1 \in \mathcal{L}_{1,1}^{(k)}} \mathbb{1}^{(k)}_{\{Q(a_1)\}} |\mathcal{Q}^{(k)}(a_1)| \mathbf{c}^{(k)}_{a'_2} + \mathbb{1}^{(k)}_{\{L_{1,1}\}} |\mathcal{L}^{(k)}_{1,1}| \mathbf{c}^{(k)}_{a'_1} \Bigg\|^2 \Bigg].
\end{aligned} \tag{40}
$$

Then, we perform the following algebraic steps to bound the second term on the right hand side of the above inequality:

$$
\begin{aligned}
&\Bigg\| \sum_{a_1 \in \mathcal{L}_{1,1}^{(k)}} \sum_{a_2 \in \mathcal{Q}^{(k)}(a_1)} \sum_{a_3 \in \mathcal{Q}^{(k)}(a_2)} \cdots \sum_{a_{|\mathcal{L}|-1} \in \mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})} \mathbb{1}^{(k)}_{\{Q(a_{|\mathcal{L}|-1})\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})| \mathbf{c}^{(k)}_{a'_{|\mathcal{L}|}} \\
&+ \sum_{a_1 \in \mathcal{L}_{1,1}^{(k)}} \sum_{a_2 \in \mathcal{Q}^{(k)}(a_1)} \sum_{a_3 \in \mathcal{Q}^{(k)}(a_2)} \cdots \sum_{a_{|\mathcal{L}|-2} \in \mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})} \mathbb{1}^{(k)}_{\{Q(a_{|\mathcal{L}|-2})\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})| \mathbf{c}^{(k)}_{a'_{|\mathcal{L}|-1}} + \cdots \\
&+ \sum_{a_1 \in \mathcal{L}_{1,1}^{(k)}} \mathbb{1}^{(k)}_{\{Q(a_1)\}} |\mathcal{Q}^{(k)}(a_1)| \mathbf{c}^{(k)}_{a'_2} + \mathbb{1}^{(k)}_{\{L_{1,1}\}} |\mathcal{L}^{(k)}_{1,1}| \mathbf{c}^{(k)}_{a'_1} \Bigg\|^2 \\
&\leq \Bigg(\Bigg\| \sum_{a_1 \in \mathcal{L}_{1,1}^{(k)}} \sum_{a_2 \in \mathcal{Q}^{(k)}(a_1)} \sum_{a_3 \in \mathcal{Q}^{(k)}(a_2)} \cdots \sum_{a_{|\mathcal{L}|-1} \in \mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})} \mathbb{1}^{(k)}_{\{Q(a_{|\mathcal{L}|-1})\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})| \mathbf{c}^{(k)}_{a'_{|\mathcal{L}|}} \Bigg\| \\
&+ \Bigg\| \sum_{a_1 \in \mathcal{L}_{1,1}^{(k)}} \sum_{a_2 \in \mathcal{Q}^{(k)}(a_1)} \sum_{a_3 \in \mathcal{Q}^{(k)}(a_2)} \cdots \sum_{a_{|\mathcal{L}|-2} \in \mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})} \mathbb{1}^{(k)}_{\{Q(a_{|\mathcal{L}|-2})\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})| \mathbf{c}^{(k)}_{a'_{|\mathcal{L}|-1}} \Bigg\| \\
&+ \cdots \\
&+ \Bigg\| \sum_{a_1 \in \mathcal{L}_{1,1}^{(k)}} \mathbb{1}^{(k)}_{\{Q(a_1)\}} |\mathcal{Q}^{(k)}(a_1)| \mathbf{c}^{(k)}_{a'_2} \Bigg\| + \Bigg\| \mathbb{1}^{(k)}_{\{L_{1,1}\}} |\mathcal{L}^{(k)}_{1,1}| \mathbf{c}^{(k)}_{a'_1} \Bigg\| \Bigg)^2 \\
&\leq \Bigg(\sum_{a_1 \in \mathcal{L}_{1,1}^{(k)}} \sum_{a_2 \in \mathcal{Q}^{(k)}(a_1)} \sum_{a_3 \in \mathcal{Q}^{(k)}(a_2)} \cdots \sum_{a_{|\mathcal{L}|-1} \in \mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})} \mathbb{1}^{(k)}_{\{Q(a_{|\mathcal{L}|-1})\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})| \Big\| \mathbf{c}^{(k)}_{a'_{|\mathcal{L}|}} \Big\| \\
&+ \sum_{a_1 \in \mathcal{L}_{1,1}^{(k)}} \sum_{a_2 \in \mathcal{Q}^{(k)}(a_1)} \sum_{a_3 \in \mathcal{Q}^{(k)}(a_2)} \cdots \sum_{a_{|\mathcal{L}|-2} \in \mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})} \mathbb{1}^{(k)}_{\{Q(a_{|\mathcal{L}|-2})\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})| \Big\| \mathbf{c}^{(k)}_{a'_{|\mathcal{L}|-1}} \Big\| \\
&+ \cdots \\
&+ \sum_{a_1 \in \mathcal{L}_{1,1}^{(k)}} \mathbb{1}^{(k)}_{\{Q(a_1)\}} |\mathcal{Q}^{(k)}(a_1)|^2 \Big\| \mathbf{c}^{(k)}_{a'_2} \Big\| + \mathbb{1}^{(k)}_{\{L_{1,1}\}} |\mathcal{L}^{(k)}_{1,1}| \Big\| \mathbf{c}^{(k)}_{a'_1} \Big\| \Bigg)^2 \\
&\overset{(a)}{\leq} \\
&\Phi \Bigg[\sum_{a_1 \in \mathcal{L}_{1,1}^{(k)}} \sum_{a_2 \in \mathcal{Q}^{(k)}(a_1)} \sum_{a_3 \in \mathcal{Q}^{(k)}(a_2)} \cdots \sum_{a_{|\mathcal{L}|-1} \in \mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})} \mathbb{1}^{(k)}_{\{Q(a_{|\mathcal{L}|-1})\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})|^2 \Big\| \mathbf{c}^{(k)}_{a'_{|\mathcal{L}|}} \Big\|^2 \\
&+ \sum_{a_1 \in \mathcal{L}_{1,1}^{(k)}} \sum_{a_2 \in \mathcal{Q}^{(k)}(a_1)} \sum_{a_3 \in \mathcal{Q}^{(k)}(a_2)} \cdots \sum_{a_{|\mathcal{L}|-2} \in \mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})} \mathbb{1}^{(k)}_{\{Q(a_{|\mathcal{L}|-2})\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})|^2 \Big\| \mathbf{c}^{(k)}_{a'_{|\mathcal{L}|-1}} \Big\|^2 \\
&+ \cdots \\
&+ \sum_{a_1 \in \mathcal{L}_{1,1}^{(k)}} \mathbb{1}^{(k)}_{\{Q(a_1)\}} |\mathcal{Q}^{(k)}(a_1)|^2 \Big\| \mathbf{c}^{(k)}_{a'_2} \Big\|^2 + \mathbb{1}^{(k)}_{\{L_{1,1}\}} |\mathcal{L}^{(k)}_{1,1}|^2 \Big\| \mathbf{c}^{(k)}_{a'_1} \Big\|^2 \Bigg],
\end{aligned} \tag{41}
$$

where the triangle inequality, e.g, for vectors $\mathbf{a}_i,\ 1 \leq i \leq n$: $\|\sum_{i=1}^{n} \mathbf{a}_i\| \leq \sum_{i=1}^{n} \|\mathbf{a}_i\|$, is applied sequentially and

$$\Phi = N_{|\mathcal{L}|-1} + N_{|\mathcal{L}|-2} + \cdots + N_1 + 1. \tag{42}$$

Also, inequality (a) in (41) is obtained by exploiting the Cauchy–Schwarz inequality,$< \mathbf{a}, \mathbf{a}' > \leq \|\mathbf{a}\| \, . \, \|\mathbf{a}'\|$, which results in the following bound, where $\mathbf{q} = [q_1, \cdots, q_b]$:

$$\left(\sum_{a=1}^{b} q_i \right)^2 = (< \mathbf{1}, \mathbf{q} >)^2 \leq b \sum_{a=1}^{b} q_i^2. \tag{43}$$

B. PART II: Finding the Consensus (local aggregation) Error in Each LUT Cluster

To further find each of the error terms in the right hand side of (41), we need to bound $\left\| \mathbf{c}_{a'_p}^{(k)} \right\|^2$, $1 \leq p \leq |\mathcal{L}|$. For notations simplicity we consider bounding $\left\| \mathbf{c}_{a'_{|\mathcal{L}|}}^{(k)} \right\|^2$ for the case where sampling is conducted from node $a'_{|\mathcal{L}|}$, $a'_{|\mathcal{L}|} \in \mathcal{C}^{(k)}$, where LUT cluster C is located in the bottom-most layer .

The evolution of nodes' parameters during D2D communications in this cluster can be described by (12) as

$$\widehat{\mathbf{W}}_C^{(k)} = \left(\mathbf{V}_C^{(k)} \right)^{\theta_C^{(k)}} \widetilde{\mathbf{W}}_C^{(k)}. \tag{44}$$

Let matrix $\overline{\mathbf{W}}_C^{(k)}$ denote the average of the vector of parameters in cluster C. This matrix can be described as

$$\overline{\mathbf{W}}_C^{(k)} = \frac{\mathbf{1}_{|\mathcal{C}^{(k)}|} \mathbf{1}_{|\mathcal{C}^{(k)}|}^{\top} \widetilde{\mathbf{W}}_C^{(k)}}{|\mathcal{C}^{(k)}|}. \tag{45}$$

We next define the local aggregation error matrix $\boldsymbol{E}_C^{(k)}$ for cluster C at the instance of global aggregation k, which satisfies the following equality:

$$\mathbf{E}_C^{(k)} = \widehat{\mathbf{W}}_C^{(k)} - \overline{\mathbf{W}}_C^{(k)}. \tag{46}$$

Note that $[\mathbf{E}_C^{(k)}]_{a'_{|\mathcal{L}|},:} = \mathbf{c}_{a'_{|\mathcal{L}|}}^{(k)}$, where $[\mathbf{E}_C^{(k)}]_{a'_{|\mathcal{L}|},:}$ denotes the row describing the parameter of node $a'_{|\mathcal{L}|}$. Note that $\mathbf{1}^{\top}\mathbf{E}_C^{(k)} = \mathbf{0}$, and thus $(\mathbf{1}\mathbf{1}^{\top})\mathbf{E}_C^{(k)} = \mathbf{0}$ and accordingly

$$\begin{aligned}
\mathbf{E}_C^{(k)} &= \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^{\top}}{|\mathcal{C}^{(k)}|} \right) \mathbf{E}_C^{(k)} = \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^{\top}}{|\mathcal{C}^{(k)}|} \right) \left(\widehat{\mathbf{W}}_C^{(k)} - \overline{\mathbf{W}}_C^{(k)} \right) \\
&= \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^{\top}}{|\mathcal{C}^{(k)}|} \right) \left(\left(\mathbf{V}_C^{(k)} \right)^{\theta_C^{(k)}} \widetilde{\mathbf{W}}_C^{(k)} - \overline{\mathbf{W}}_C^{(k)} \right) = \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^{\top}}{|\mathcal{C}^{(k)}|} \right) \left(\left(\mathbf{V}_C^{(k)} \right)^{\theta_C^{(k)}} \widetilde{\mathbf{W}}_C^{(k)} - \left(\mathbf{V}_C^{(k)} \right)^{\theta_C^{(k)}} \overline{\mathbf{W}}_C^{(k)} \right) \\
&= \left(\left(\mathbf{V}_C^{(k)} \right)^{\theta_C^{(k)}} - \frac{\mathbf{1}\mathbf{1}^{\top}}{|\mathcal{C}^{(k)}|} \right) \left(\widetilde{\mathbf{W}}_C^{(k)} - \overline{\mathbf{W}}_C^{(k)} \right),
\end{aligned} \tag{47}$$

where $\mathbf{I}$ denotes the identity matrix. In the above equalities we have used the facts that (i) $\left(\mathbf{V}_C^{(k)} \right)^{\theta_C^{(k)}} \overline{\mathbf{W}}_C^{(k)} = \overline{\mathbf{W}}_C^{(k)}$ since performing consensus on averaged matrix does not change the resulting parameters, and (ii) $\frac{\mathbf{1}\mathbf{1}^{\top}}{|\mathcal{C}^{(k)}|} \left(\mathbf{V}_C^{(k)} \right)^{\theta_C^{(k)}} = \frac{\mathbf{1}\mathbf{1}^{\top}}{|\mathcal{C}^{(k)}|}$ according to Assumption 2 since $\left(\mathbf{V}_C^{(k)} \right)^{\theta_C^{(k)}}$ is also double stochastic.

Using the above properties, we finally bound $\|\mathbf{c}_{a'_{|\mathcal{L}|}}^{(k)}\|$ as follows:

$$\begin{aligned}
\left\| \mathbf{c}_{a'_{|\mathcal{L}|}}^{(k)} \right\|^2 &\leq \text{trace} \left((\mathbf{E}_C^{(k)})^{\top} \mathbf{E}_C^{(k)} \right) = \text{trace} \left(\left(\widetilde{\mathbf{W}}_C^{(k)} - \overline{\mathbf{W}}_C^{(k)} \right)^{\top} \left(\left(\mathbf{V}_C^{(k)} \right)^{\theta_C^{(k)}} - \frac{\mathbf{1}\mathbf{1}^{\top}}{|\mathcal{C}^{(k)}|} \right)^2 \left(\widetilde{\mathbf{W}}_C^{(k)} - \overline{\mathbf{W}}_C^{(k)} \right) \right) \\
&\leq (\lambda_C^{(k)})^{2\theta_C^{(k)}} \sum_{q \in \mathcal{C}^{(k)}} \left\| \widetilde{\mathbf{w}}_q^{(k)} - \overline{\mathbf{w}}_C^{(k)} \right\|^2 \leq (\lambda_C^{(k)})^{2\theta_C^{(k)}} \frac{1}{|\mathcal{C}^{(k)}|} \sum_{q,q' \in \mathcal{C}^{(k)}} \left\| \widetilde{\mathbf{w}}_q^{(k)} - \overline{\mathbf{w}}_{q'}^{(k)} \right\|^2 \\
&\leq (\lambda_C^{(k)})^{2\theta_C^{(k)}} |\mathcal{C}^{(k)}| \max_{q,q' \in \mathcal{C}^{(k)}} \left\| \widetilde{\mathbf{w}}_q^{(k)} - \overline{\mathbf{w}}_{q'}^{(k)} \right\|^2 \leq (\lambda_C^{(k)})^{2\theta_C^{(k)}} |\mathcal{C}^{(k)}| \left(\Upsilon_C^{(k)} \right)^2,
\end{aligned} \tag{48}$$

where $\overline{\mathbf{w}}_C^{(k)}$ denotes the vector of average of parameters inside the cluster and we used the fact that $\left(\mathbf{V}_C^{(k)} \right)^{\theta_C^{(k)}} - \frac{\mathbf{1}^{\top}}{|\mathcal{C}^{(k)}|} =$

$\left(\mathbf{V}_C^{(k)}\right)^{\theta_C^{(k)}}\left(\mathbf{I}-\frac{\mathbf{1}\mathbf{1}^\top}{|\mathcal{C}^{(k)}|}\right)=\left(\mathbf{V}_C^{(k)}\right)^{\theta_C^{(k)}}\left(\mathbf{I}-\frac{\mathbf{1}\mathbf{1}^\top}{|\mathcal{C}^{(k)}|}\right)^{\theta_C^{(k)}}=\left(\mathbf{V}_C^{(k)}-\frac{\mathbf{1}\mathbf{1}^\top}{|\mathcal{C}^{(k)}|}\right)^{\theta_C^{(k)}}$ (note that $\left(\mathbf{I}-\frac{\mathbf{1}\mathbf{1}^\top}{|\mathcal{C}^{(k)}|}\right)$ is a projection matrix) is a real symmetric matrix.

The above mentioned proof can be generalized to every cluster with slight modifications, which results in

$$\left\|\mathbf{c}_{a'_p}^{(k)}\right\|^2\leq(\lambda_C)^{2\theta_C^{(k)}}|\mathcal{C}^{(k)}|\left(\Upsilon_C^{(k)}\right)^2,\ a'_p\in\mathcal{C}. \tag{49}$$

C. PART III: Obtaining the Final Convergence Bound

Replacing the above inequality in (41) combined with (40) gives us

$$\begin{aligned}
&F(\mathbf{w}^{(k)})-F(\mathbf{w}^{(k-1)})\leq\frac{-\mu}{\eta}\Big(F(\mathbf{w}^{(k-1)})-F(\mathbf{w}^*)\Big)+\\
&\frac{\eta\Phi}{2D^2}\Bigg[\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\cdots\sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})}\mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-1})\right\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})|^3\left(\lambda^{(k)}_{Q(a_{|\mathcal{L}|-1})}\right)^{2\theta^{(k)}_{Q(a_{|\mathcal{L}|-1})}}\left(\Upsilon^{(k)}_{Q(a_{|\mathcal{L}|-1})}\right)^2\\
&+\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\cdots\sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})}\mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-2})\right\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})|^3\left(\lambda^{(k)}_{Q(a_{|\mathcal{L}|-2})}\right)^{2\theta^{(k)}_{Q(a_{|\mathcal{L}|-2})}}\left(\Upsilon^{(k)}_{Q(a_{|\mathcal{L}|-2})}\right)^2\\
&+\cdots+\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\mathbb{1}^{(k)}_{\left\{Q(a_1)\right\}}|\mathcal{Q}^{(k)}(a_1)|^3\left(\lambda^{(k)}_{Q(a_1)}\right)^{2\theta^{(k)}_{Q(a_1)}}\left(\Upsilon^{(k)}_{Q(a_1)}\right)^2+\mathbb{1}^{(k)}_{\left\{L_{1,1}\right\}}|\mathcal{L}^{(k)}_{1,1}|^3\left(\lambda^{(k)}_{L_{1,1}}\right)^{2\theta^{(k)}_{L_{1,1}}}\left(\Upsilon^{(k)}_{L_{1,1}}\right)^2\Bigg].
\end{aligned} \tag{50}$$

Adding $F(\mathbf{w}^{(k-1)})$ to both hands sides and subtracting $F(\mathbf{w}^*)$ from both hand sides, we get

$$\begin{aligned}
&F(\mathbf{w}^{(k)})-F(\mathbf{w}^*)\leq(1-\frac{\mu}{\eta})\underbrace{\Big(F(\mathbf{w}^{(k-1)})-F(\mathbf{w}^*)\Big)}_{(a)}+\\
&\frac{\eta\Phi}{2D^2}\Bigg[\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\cdots\sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})}\mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-1})\right\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})|^3\left(\lambda^{(k)}_{Q(a_{|\mathcal{L}|-1})}\right)^{2\theta^{(k)}_{Q(a_{|\mathcal{L}|-1})}}\left(\Upsilon^{(k)}_{Q(a_{|\mathcal{L}|-1})}\right)^2\\
&+\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\cdots\sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})}\mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-2})\right\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})|^3\left(\lambda^{(k)}_{Q(a_{|\mathcal{L}|-2})}\right)^{2\theta^{(k)}_{Q(a_{|\mathcal{L}|-2})}}\left(\Upsilon^{(k)}_{Q(a_{|\mathcal{L}|-2})}\right)^2\\
&+\cdots+\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\mathbb{1}^{(k)}_{\left\{Q(a_1)\right\}}|\mathcal{Q}^{(k)}(a_1)|^3\left(\lambda^{(k)}_{Q(a_1)}\right)^{2\theta^{(k)}_{Q(a_1)}}\left(\Upsilon^{(k)}_{Q(a_1)}\right)^2+\mathbb{1}^{(k)}_{\left\{L_{1,1}\right\}}|\mathcal{L}^{(k)}_{1,1}|^3\left(\lambda^{(k)}_{L_{1,1}}\right)^{2\theta^{(k)}_{L_{1,1}}}\left(\Upsilon^{(k)}_{L_{1,1}}\right)^2\Bigg].
\end{aligned} \tag{51}$$

Expanding term (a) on the right hand side of the inequality in a recursive manner leads to the theorem result. ■

APPENDIX B
PROOF OF PROPOSITION 1

Consider the bound on the number of D2D that is given in the proposition statement. For $L_{j,i}$, if $\sigma_j\leq|\mathcal{L}^{(k)}_{j,i}|^3\left(\Upsilon^{(k)}_{L_{j,i}}\right)^2$, $\forall i$, the proposed number of D2D guarantees $\theta^{(k)}_{L_{j,i}}\geq\frac{\log(\sigma_j)-2\log\left(\left|\mathcal{L}^{(k)}_{j,i}\right|^{\frac{3}{2}}\Upsilon^{(k)}_{L_{j,i}}\right)}{2\log\left(\lambda^{(k)}_{L_{j,i}}\right)}$, which results in

$$\begin{aligned}
\theta^{(k)}_{L_{j,i}}&\geq\frac{\log(\sigma_j)-2\log\left(\left|\mathcal{L}^{(k)}_{j,i}\right|^{\frac{3}{2}}\Upsilon^{(k)}_{L_{j,i}}\right)}{2\log\left(\lambda^{(k)}_{L_{j,i}}\right)}\\
\Rightarrow\theta^{(k)}_{L_{j,i}}&\geq\frac{1}{2}\frac{\log\left(\frac{\sigma_j}{|\mathcal{L}^{(k)}_{j,i}|^3\left(\Upsilon^{(k)}_{L_{j,i}}\right)^2}\right)}{\log\left(\lambda^{(k)}_{L_{j,i}}\right)}\\
\Rightarrow\left(\lambda^{(k)}_{L_{j,i}}\right)^{2\theta^{(k)}_{L_{j,i}}}&\leq\frac{\sigma_j}{|\mathcal{L}^{(k)}_{j,i}|^3\left(\Upsilon^{(k)}_{L_{j,i}}\right)^2},\ \forall k,
\end{aligned} \tag{52}$$

where the last inequality is due to the facts that $\frac{\log a}{\log b} = \log_a^b$, $a^{\log_a^b} = b$, and $\lambda_{L_{j,i}}^{(k)} < 1$. Also, for cluster $L_{j,i}$, if $\sigma_j \geq |\mathcal{L}_{j,i}^{(k)}|^3 \left(\Upsilon_{L_{j,i}}^{(k)}\right)^2$, any $\theta_{L_{j,i}}^{(k)} \geq 0$ ensures $\sigma_j \geq |\mathcal{L}_{j,i}^{(k)}|^3 \left(\Upsilon_{L_{j,i}}^{(k)}\right)^2 \left(\lambda_{L_{j,i}}^{(k)}\right)^{2\theta_{L_{j,i}}^{(k)}}$, $\forall k$. Replacing the above result in (14), we get

$$\begin{aligned}
&F(\mathbf{w}^{(k-1)}) - F(\mathbf{w}^*) \leq \frac{\eta\Phi}{2D^2} \sum_{t=0}^{k-1} \left(\frac{\eta-\mu}{\eta}\right)^t \Bigg[\\
&\sum_{a_1 \in \mathcal{L}_{1,1}^{(k)}} \sum_{a_2 \in \mathcal{Q}^{(k)}(a_1)} \cdots \sum_{a_{|\mathcal{L}|-1} \in \mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})} \mathbb{1}_{\left\{\mathcal{Q}(a_{|\mathcal{L}|-1})\right\}}^{(k-t)} \sigma_{|\mathcal{L}|} + \\
&\sum_{a_1 \in \mathcal{L}_{1,1}^{(k)}} \sum_{a_2 \in \mathcal{Q}^{(k)}(a_1)} \cdots \sum_{a_{|\mathcal{L}|-2} \in \mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})} \mathbb{1}_{\left\{\mathcal{Q}(a_{|\mathcal{L}|-2})\right\}}^{(k-t)} \sigma_{|\mathcal{L}|-1} + \cdots \\
&+ \sum_{a_1 \in \mathcal{L}_{1,1}^{(k)}} \mathbb{1}_{\{\mathcal{Q}(a_1)\}}^{(k-t)} \sigma_2 + \mathbb{1}_{\{L_{1,1}\}}^{(k-t)} \sigma_1 \Bigg] + \left(\frac{\eta-\mu}{\eta}\right)^k \Big(F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*)\Big) \\
&\leq \frac{\eta\Phi}{2D^2} \sum_{t=0}^{k-1} \left(\frac{\eta-\mu}{\eta}\right)^t \Big[N_{|\mathcal{L}|-1}\sigma_{|\mathcal{L}|} + N_{|\mathcal{L}|-2}\sigma_{|\mathcal{L}|-1} + \cdots \\
&+ N_1\sigma_2 + N_0\sigma_1 \Big] + \left(\frac{\eta-\mu}{\eta}\right)^k \Big(F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*)\Big).
\end{aligned} \tag{53}$$

Taking the limit with respect to k, we get

$$\lim_{k\to\infty} F(\mathbf{w}^{(k-1)}) - F(\mathbf{w}^*) \leq \frac{\eta\Phi}{2D^2} \left(\sum_{j=0}^{|\mathcal{L}|-1} \sigma_{j+1} N_j\right) \frac{1}{1-(\frac{\eta-\mu}{\eta})}, \tag{54}$$

which concludes the proof.

Appendix C
Proof of Proposition 2

Consider the per-iteration decrease of the objective function given by (51). Following a similar procedure as Appendix B, given the proposed number of D2D rounds in the proposition statement, we get

$$F(\mathbf{w}^{(k)}) - F(\mathbf{w}^*) \leq (1-\frac{\mu}{\eta}) \Big(F(\mathbf{w}^{(k-1)}) - F(\mathbf{w}^*)\Big) + \frac{\eta\Phi}{2D^2} \left[\sum_{j=0}^{|\mathcal{L}|-1} \sigma_{j+1}^{(k)} N_j\right]. \tag{55}$$

Using the fact that $\nabla F(\mathbf{w}^*) = 0$ combined with η-smoothness of F, we get

$$\left\|\nabla F(\mathbf{w}^{(k-1)})\right\| = \left\|\nabla F(\mathbf{w}^{(k-1)}) - \nabla F(\mathbf{w}^*)\right\| \leq \eta \|\mathbf{w}^{(k-1)} - \mathbf{w}^*\|. \tag{56}$$

Also, it is straightforward to verify that strong convexity of F, expressed in Assumption 1, implies the following inequality:

$$\mu/2 \left\|\mathbf{w}^{(k-1)} - \mathbf{w}^*\right\|^2 \leq F(\mathbf{w}^{(k-1)}) - F(\mathbf{w}^*). \tag{57}$$

Combining the above results with the condition given in the proposition statement, i.e., (19), we get

$$\begin{aligned}
\sum_{j=0}^{|\mathcal{L}|-1} \sigma_{j+1}^{(k)} N_j &\leq \frac{D^2\mu(\mu-\delta\eta)}{\eta^4\Phi} \left\|\nabla F(\mathbf{w}^{(k-1)})\right\|^2 \\
&\leq \frac{D^2\mu(\mu-\delta\eta)}{\eta^2\Phi} \left\|\mathbf{w}^{(k-1)} - \mathbf{w}^*\right\|^2 \\
&\leq \frac{2D^2(\mu-\delta\eta)}{\eta^2\Phi} \Big(F(\mathbf{w}^{(k-1)}) - F(\mathbf{w}^*)\Big)
\end{aligned} \tag{58}$$

By replacing the above inequality in (55) we obtain

$$F(\mathbf{w}^{(k)}) - F(\mathbf{w}^*) \leq (1-\mu/\eta) \Big(F(\mathbf{w}^{(k-1)}) - F(\mathbf{w}^*)\Big) + (\mu/\eta - \delta) \Big(F(\mathbf{w}^{(k-1)}) - F(\mathbf{w}^*)\Big), \tag{59}$$

which readily leads to the proposition result.

APPENDIX D
PROOF OF COROLLARY 1

Regarding the first condition, at global iteration κ, using the number of consensus given in the corollary statement, according to (53), we have

$$\begin{aligned} F(\mathbf{w}^{(\kappa)}) - F(\mathbf{w}^*) &\leq \frac{\eta\Phi}{2D^2} \sum_{j=0}^{|\mathcal{L}|-1} \sigma_{j+1} N_j \frac{1-\left(1-\frac{\mu}{\eta}\right)^{\kappa}}{\mu/\eta} + \left(\frac{\eta-\mu}{\eta}\right)^{\kappa} \left(F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*)\right) \\ &= \left(1-\frac{\mu}{\eta}\right)^{\kappa} \left(F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*) - \frac{\eta^2\Phi}{2\mu D^2} \sum_{j=0}^{|\mathcal{L}|-1} \sigma_{j+1} N_j\right) + \frac{\eta^2\Phi}{2\mu D^2} \sum_{j=0}^{|\mathcal{L}|-1} \sigma_{j+1} N_j. \end{aligned} \tag{60}$$

Thus to satisfy the accuracy requirement, it is sufficient to have

$$\left(1-\frac{\mu}{\eta}\right)^{\kappa} \left(F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*) - \frac{\eta^2\Phi}{2\mu D^2} \sum_{j=0}^{|\mathcal{L}|-1} \sigma_{j+1} N_j\right) + \frac{\eta^2\Phi}{2\mu D^2} \sum_{j=0}^{|\mathcal{L}|-1} \sigma_{j+1} N_j \leq \epsilon. \tag{61}$$

Performing some algebraic steps leads to (23).

Regarding the second condition, given the number of D2D rounds stated in the proposition statement, we first recursively expand the right hand side of (59) to get

$$F(\mathbf{w}^{(\kappa)}) - F(\mathbf{w}^*) \leq (1-\delta)^{\kappa} \left(F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*)\right). \tag{62}$$

Thus, to satisfy the desired accuracy, it is sufficient to have

$$(1-\delta)^{\kappa}[F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*)] \leq \epsilon, \tag{63}$$

which readily leads to (24). Note that the criterion given in the corollary statement for ϵ guarantees that: $0 < \delta \leq \mu/\eta$.

APPENDIX E
PROOF OF COROLLARY 2

Regarding the first condition, upon using the number of D2D rounds described in the corollary statement, we get (61), which can be written as

$$\left(1-\frac{\mu}{\eta}\right)^{\kappa} \leq \frac{\epsilon - \frac{\eta^2\Phi}{2\mu D^2} \sum_{j=0}^{|\mathcal{L}|-1} \sigma_{j+1} N_j}{F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*) - \frac{\eta^2\Phi}{2\mu D^2} \sum_{j=0}^{|\mathcal{L}|-1} \sigma_{j+1} N_j}. \tag{64}$$

To obtain κ, we need to take the logarithm with base $1-\mu/\eta$, where $0 < 1-\mu/\eta < 1$. Using the characteristic of the logarithm upon having a positive base less than one, we get

$$\kappa \geq \log_{1-\mu/\eta}^{\left(\epsilon - \frac{\eta^2\Phi}{2\mu D^2} \sum_{j=0}^{|\mathcal{L}|-1} \sigma_{j+1} N_j\right) \left(F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*) - \frac{\eta^2\Phi}{2\mu D^2} \sum_{j=0}^{|\mathcal{L}|-1} \sigma_{j+1} N_j\right)^{-1}}, \tag{65}$$

which can be written as (21).

Regarding the second condition, upon using the number of D2D rounds described in the corollary statement, we get (63). To obtain κ, we take the logarithm with base $1-\delta$ from both hand sides of the equation, using the fact that $0 < 1-\delta < 1$ and the characteristic of the logarithm upon having a positive base less than one, we get

$$\kappa \geq \log_{1-\delta}^{\left(\epsilon / \left(F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*)\right)\right)}, \tag{66}$$

which can be written as (22).

APPENDIX F
PROOF OF PROPOSITION 3

Upon sharing the gradients, the nodes in the bottom layer share their scaled gradients (multiplying their gradients by their number of data points), while the rest of the procedure, i.e., traversing of the gradients over the hierarchy, is the same as sharing the parameters. For parent node a_p, let a'_{p+1} denote the corresponding sampled node, $\forall p$, e.g., in the following nested

sums $a'_{|\mathcal{L}|}$ denotes the sampled node in the last layer by parent node $a_{|\mathcal{L}|-1}$ in its above layer. Let $\hat{\mathbf{g}}_{a'_1}^{(k)}$ denote the sampled value by the main server at global iteration k. It can be verified that we have

$$\begin{aligned}
\hat{\mathbf{g}}_{a'_1}^{(k)} = & \frac{\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}|\mathcal{D}_{a_{|\mathcal{L}|}}|\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}_{a_{|\mathcal{L}|}}^{(k-1)})}{|\mathcal{L}_{1,1}^{(k)}|} \\
& + \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})}\frac{\mathbb{1}_{\{Q(a_{|\mathcal{L}|-1})\}}^{(k)}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})|\mathbf{c}_{a'_{|\mathcal{L}|}}^{(k)}}{|\mathcal{L}_{1,1}^{(k)}|} \\
& + \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})}\frac{\mathbb{1}_{\{Q(a_{|\mathcal{L}|-2})\}}^{(k)}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})|\mathbf{c}_{a'_{|\mathcal{L}|-1}}^{(k)}}{|\mathcal{L}_{1,1}^{(k)}|}+ \\
& \vdots \\
& + \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\frac{\mathbb{1}_{\{Q(a_1)\}}^{(k)}|\mathcal{Q}^{(k)}(a_1)|\mathbf{c}_{a'_2}^{(k)}}{|\mathcal{L}_{1,1}^{(k)}|}+\mathbb{1}_{\{L_{1,1}\}}^{(k)}\mathbf{c}_{a'_1}^{(k)}.
\end{aligned} \tag{67}$$

The main server then uses this vector as the estimation of global gradient and builds the parameter vector for the next iteration as follows (note that although the root only receives the gradients, it has the knowledge of the previous parameters that it broadcast, i.e., $\mathbf{w}^{(k-1)}$):

$$\begin{aligned}
\widehat{\mathbf{w}}_{a'_1}^{(k)} = & D\frac{\mathbf{w}^{(k-1)}}{|\mathcal{L}_{1,1}^{(k)}|}- \\
& \beta_{k-1}\Bigg[\frac{\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}|\mathcal{D}_{a_{|\mathcal{L}|}}|\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}_{a_{|\mathcal{L}|}}^{(k-1)})}{|\mathcal{L}_{1,1}^{(k)}|} \\
& + \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})}\frac{\mathbb{1}_{\{Q(a_{|\mathcal{L}|-1})\}}^{(k)}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})|\mathbf{c}_{a'_{|\mathcal{L}|}}^{(k)}}{|\mathcal{L}_{1,1}^{(k)}|} \\
& + \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})}\frac{\mathbb{1}_{\{Q(a_{|\mathcal{L}|-2})\}}^{(k)}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})|\mathbf{c}_{a'_{|\mathcal{L}|-1}}^{(k)}}{|\mathcal{L}_{1,1}^{(k)}|}+ \\
& \vdots \\
& + \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\frac{\mathbb{1}_{\{Q(a_1)\}}^{(k)}|\mathcal{Q}^{(k)}(a_1)|\mathbf{c}_{a'_2}^{(k)}}{|\mathcal{L}_{1,1}^{(k)}|}+\mathbb{1}_{\{L_{1,1}\}}^{(k)}\mathbf{c}_{a'_1}^{(k)}\Bigg],
\end{aligned} \tag{68}$$

which is used to obtain the next global parameter (due to the existence of the indicator function in the last term of the above expression, the following expression holds regardless of the operating mode of the cluster at layer L_1):

$$\mathbf{w}^{(k)} = \frac{|\mathcal{L}_{1,1}^{(k)}|\widehat{\mathbf{w}}_{a'_1}^{(k)}}{D}. \tag{69}$$

According to (1), it can be verified that

$$\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}|\mathcal{D}_{a_{|\mathcal{L}|}}|\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}_{a_{|\mathcal{L}|}}^{(k-1)}) = D\nabla F(\mathbf{w}^{(k-1)}). \tag{70}$$

Replacing the above equation in (68) and performing the update given by (69), we get

$$\begin{aligned}\mathbf{w}^{(k)} = \mathbf{w}^{(k-1)} - \beta_{k-1}\Bigg[\nabla F(\mathbf{w}^{(k-1)}) + \frac{1}{D}\Bigg(&\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\cdots\sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})}\mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-1})\right\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})|\mathbf{c}^{(k)}_{a'_{|\mathcal{L}|}}\\ &+\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\cdots\sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})}\mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-2})\right\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})|\mathbf{c}^{(k)}_{a'_{|\mathcal{L}|-1}}+\\ &\vdots\\ &+\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\mathbb{1}^{(k)}_{\{Q(a_1)\}}|\mathcal{Q}^{(k)}(a_1)|\mathbf{c}^{(k)}_{a'_2}+\mathbb{1}^{(k)}_{\{L_{1,1}\}}|\mathcal{L}^{(k)}_{1,1}|\mathbf{c}^{(k)}_{a'_1}\Bigg)\Bigg].\end{aligned} \tag{71}$$

Using the above equality in (31), we have

$$\begin{aligned}F(\mathbf{w}^{(k)}) &\leq F(\mathbf{w}^{(k-1)}) - \beta_{k-1}\left(\nabla F(\mathbf{w}^{(k-1)}) + \mathbf{c}^{(k)}\right)^\top \nabla F(\mathbf{w}^{(k-1)}) + \frac{\eta}{2}\beta_{k-1}^2\|\nabla F(\mathbf{w}^{(k-1)}) + \mathbf{c}^{(k)}\|^2\\ &= F(\mathbf{w}^{(k-1)}) - \beta_{k-1}\left\|\nabla F(\mathbf{w}^{(k-1)})\right\|^2 - \beta_{k-1}\left(\mathbf{c}^{(k)}\right)^\top \nabla F(\mathbf{w}^{(k-1)}) + \frac{\eta}{2}\beta_{k-1}^2\|\nabla F(\mathbf{w}^{(k-1)}) + \mathbf{c}^{(k)}\|^2\\ &= F(\mathbf{w}^{(k-1)}) - \beta_{k-1}\left\|\nabla F(\mathbf{w}^{(k-1)})\right\|^2 - \beta_{k-1}\left(\mathbf{c}^{(k)}\right)^\top \nabla F(\mathbf{w}^{(k-1)}) + \frac{\eta\beta_{k-1}^2}{2}\|\nabla F(\mathbf{w}^{(k-1)})\|^2\\ &+ \beta_{k-1}^2\eta\left(\nabla F(\mathbf{w}^{(k-1)})^\top\mathbf{c}^{(k)}\right) + \frac{\eta\beta_{k-1}^2}{2}\left\|\mathbf{c}^{(k)}\right\|^2,\end{aligned} \tag{72}$$

where

$$\begin{aligned}\mathbf{c}^{(k)} \triangleq \frac{1}{D}\Bigg(&\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})}\mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-1})\right\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})|\mathbf{c}^{(k)}_{a'_{|\mathcal{L}|}}\\ &+\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})}\mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-2})\right\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})|\mathbf{c}^{(k)}_{a'_{|\mathcal{L}|-1}}+\\ &\vdots\\ &+\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\mathbb{1}^{(k)}_{\{Q(a_1)\}}|\mathcal{Q}^{(k)}(a_1)|\mathbf{c}^{(k)}_{a'_2}+\mathbb{1}^{(k)}_{\{L_{1,1}\}}|\mathcal{L}^{(k)}_{1,1}|\mathbf{c}^{(k)}_{a'_1}\Bigg).\end{aligned} \tag{73}$$

Taking the expectation from both hand sides (with respect to the consensus errors) and using the fact that upon using the consensus method, when one node is sampled uniformly at random we have:[5] $\mathbb{E}[\mathbf{c}^{(k)}_{a'_p}] = \mathbf{0}$, $\forall p$. This implies $\mathbb{E}[\mathbf{c}^{(k)}] = \mathbf{0}$, $\forall k$, replacing which in (72) gives us

$$\mathbb{E}[F(\mathbf{w}^{(k)})] \leq F(\mathbf{w}^{(k-1)}) - (1-\frac{\eta\beta_{k-1}}{2})\beta_{k-1}\left\|\nabla F(\mathbf{w}^{(k-1)})\right\|^2 + \frac{\eta\beta_{k-1}^2}{2}E[\|\mathbf{c}^{(k)}\|^2]. \tag{74}$$

Using the fact that $\beta_0 \leq 1/\eta$, we get $\beta_k \leq 1/\eta$, and thus $1-\eta\beta_k/2 \geq 1/2$, $\forall k$. Using this in the above inequality gives us

$$\mathbb{E}[F(\mathbf{w}^{(k)})] \leq F(\mathbf{w}^{(k-1)}) - \frac{\beta_{k-1}}{2}\left\|\nabla F(\mathbf{w}^{(k-1)})\right\|^2 + \frac{\eta\beta_{k-1}^2}{2}E[\|\mathbf{c}^{(k)}\|^2]. \tag{75}$$

Using the strong convexity, we get Polyak-Lojasiewicz inequality [55] in the following form: $\|\nabla F(\mathbf{w}^{(k-1)})\|^2 \geq 2\mu[F(\mathbf{w}^{(k-1)}) - F(\mathbf{w}^*)]$, using which in the above inequality yields

$$\mathbb{E}[F(\mathbf{w}^{(k)})] \leq F(\mathbf{w}^{(k-1)}) - \beta_{k-1}\mu[F(\mathbf{w}^{(k-1)}) - F(\mathbf{w}^*)] + \frac{\eta\beta_{k-1}^2}{2}E[\|\mathbf{c}^{(k)}\|^2], \tag{76}$$

or, equivalently

$$\mathbb{E}[F(\mathbf{w}^{(k)})] - F(\mathbf{w}^*) \leq (1-\beta_{k-1}\mu)[F(\mathbf{w}^{(k-1)}) - F(\mathbf{w}^*)] + \frac{\eta\beta_{k-1}^2}{2}E[\|\mathbf{c}^{(k)}\|^2]. \tag{77}$$

Taking total expectation, with respect to all the consensus errors until iteration k, from both hand sides results in

$$\mathbb{E}[F(\mathbf{w}^{(k)}) - F(\mathbf{w}^*)] \leq (1-\beta_{k-1}\mu)\mathbb{E}[F(\mathbf{w}^{(k-1)}) - F(\mathbf{w}^*)] + \frac{\eta\beta_{k-1}^2}{2}E[\|\mathbf{c}^{(k)}\|^2]. \tag{78}$$

[5]Assume a set of n numbers denoted by $x_1, \cdots, x_n$ with mean $\bar{x}$. Assume that X denotes a random variable with probability mass function $p(X = x_i) = \frac{1}{n}$, $1 \leq i \leq n$. It is straightforward to verify that $E(X - \bar{x}) = 0$.

We continue the proof by carrying out an induction. The proposition result trivially holds for iteration 0. Assume that the result holds for iteration k, i.e., $\mathbb{E}[F(\mathbf{w}^{(k)}) - F(\mathbf{w}^*)] \leq \frac{\Gamma}{k+\lambda}$. We aim to show that the result also holds for iteration $k+1$. Using (78), we get

$$\mathbb{E}[F(\mathbf{w}^{(k+1)}) - F(\mathbf{w}^*)] \leq (1-\beta_k\mu)\mathbb{E}[F(\mathbf{w}^{(k)}) - F(\mathbf{w}^*)] + \frac{\eta\beta_k^2}{2}E[\|\mathbf{c}^{(k+1)}\|^2], \tag{79}$$

which results in

$$\begin{aligned}\mathbb{E}[F(\mathbf{w}^{(k+1)}) - F(\mathbf{w}^*)] &\leq (1-\frac{\alpha}{k+\lambda}\mu)\frac{\Gamma}{k+\lambda} + \frac{\eta\alpha^2}{2(k+\lambda)^2}\mathbb{E}[\|\mathbf{c}^{(k+1)}\|^2] \\ &= \left(\frac{k+\lambda-\alpha\mu}{(k+\lambda)^2}\right)\Gamma + \frac{\eta\alpha^2}{2(k+\lambda)^2}\mathbb{E}[\|\mathbf{c}^{(k+1)}\|^2] \\ &= \left(\frac{k+\lambda-1}{(k+\lambda)^2}\right)\Gamma - \frac{\alpha\mu-1}{(k+\lambda)^2}\Gamma + \frac{\eta\alpha^2}{2(k+\lambda)^2}\mathbb{E}[\|\mathbf{c}^{(k+1)}\|^2].\end{aligned} \tag{80}$$

Note that using a similar method as Appendix A, we can get[6]

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{c}^{(k+1)}\right\|^2\right] \leq \frac{\Phi}{D^2}\Bigg[& \\ &\sum_{a_1\in\mathcal{L}_{1,1}^{(k+1)}}\sum_{a_2\in\mathcal{Q}^{(k+1)}(a_1)}\cdots\sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k+1)}(a_{|\mathcal{L}|-2})}\mathbb{1}^{(k+1)}_{\left\{Q(a_{|\mathcal{L}|-1})\right\}}|\mathcal{Q}^{(k+1)}(a_{|\mathcal{L}|-1})|^3\left(\lambda^{(k+1)}_{Q(a_{|\mathcal{L}|-1})}\right)^{2\theta^{(k+1)}_{Q(a_{|\mathcal{L}|-1})}}\left(\Upsilon^{(k+1)}_{Q(a_{|\mathcal{L}|-1})}\right)^2 \\ &+\sum_{a_1\in\mathcal{L}_{1,1}^{(k+1)}}\sum_{a_2\in\mathcal{Q}^{(k+1)}(a_1)}\cdots\sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k+1)}(a_{|\mathcal{L}|-3})}\mathbb{1}^{(k+1)}_{\left\{Q(a_{|\mathcal{L}|-2})\right\}}|\mathcal{Q}^{(k+1)}(a_{|\mathcal{L}|-2})|^3\left(\lambda^{(k+1)}_{Q(a_{|\mathcal{L}|-2})}\right)^{2\theta^{(k+1)}_{Q(a_{|\mathcal{L}|-2})}}\left(\Upsilon^{(k+1)}_{Q(a_{|\mathcal{L}|-2})}\right)^2 \\ &+\cdots+\sum_{a_1\in\mathcal{L}_{1,1}^{(k+1)}}\mathbb{1}^{(k+1)}_{\left\{Q^{(k+1)}(a_1)\right\}}|\mathcal{Q}^{(k+1)}(a_1)|^3\left(\lambda^{(k+1)}_{Q(a_1)}\right)^{2\theta^{(k+1)}_{Q(a_1)}}\left(\Upsilon^{(k+1)}_{Q(a_1)}\right)^2+\mathbb{1}^{(k+1)}_{\{L_{1,1}\}}|\mathcal{L}^{(k+1)}_{1,1}|^3\left(\lambda^{(k+1)}_{L_{1,1}}\right)^{2\theta^{(k+1)}_{L_{1,1}}}\left(\Upsilon^{(k+1)}_{L_{1,1}}\right)^2\Bigg].\end{aligned} \tag{81}$$

Using the number of D2D rounds given in the proposition, similar to the approach taken in Appendix B it can be verified that $E[\|\mathbf{c}^{(k+1)}\|^2] \leq C = \frac{\Phi}{D^2}\sum_{j=0}^{|\mathcal{L}|-1}N_j\sigma_{j+1}$, $\forall k$. Using this and the definition of Γ in (26), we get: $\Gamma \geq \frac{\eta\alpha^2 C}{2(\alpha\mu-1)}$, $\forall k$. Using this result in the last line of (80), we get

$$\mathbb{E}[F(\mathbf{w}^{(k+1)}) - F(\mathbf{w}^*)] \leq \left(\frac{k+\lambda-1}{(k+\lambda)^2}\right)\Gamma. \tag{82}$$

Note that since $k+\lambda > 1$, we have $(k+\lambda)^2 \geq (k+\lambda-1)(k+\lambda+1)$. Using this fact in (82), we obtain

$$\mathbb{E}[F(\mathbf{w}^{(k+1)}) - F(\mathbf{w}^*)] \leq \left(\frac{1}{k+\lambda+1}\right)\Gamma, \tag{83}$$

which completes the induction and thus the proof.

[6]Note that if for every realization of random variable X, inequality $\|X\|^2 < y$ holds, then we get: $\mathbb{E}[\|X\|^2] < y$.

$$\begin{aligned}F(\mathbf{w}^{(k)}) - F(\mathbf{w}^*) \leq &\left[\prod_{l=1}^{k}\left(1-\frac{\mu}{\eta}+8\frac{c_2}{D^2}\left(D-D_s^{(l)}\right)^2\right)\right]\left(F(\mathbf{w}^{(0)})-F(\mathbf{w}^*)\right)+\left(\sum_{t=1}^{k}\left[\prod_{l=t+1}^{k}\left(1-\frac{\mu}{\eta}+8\frac{c_2}{D^2}\left(D-D_s^{(l)}\right)^2\right)\right]\right. \\ &\left(\frac{\eta\Phi}{\left(D_s^{(t)}\right)^2}\Big[\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\cdots\sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})}\mathbb{1}^{(t)}_{\left\{\mathcal{Q}(a_{|\mathcal{L}|-1})\right\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})|^3\left(\lambda^{(t)}_{\mathcal{Q}(a_{|\mathcal{L}|-1})}\right)^{2\theta^{(t)}_{\mathcal{Q}(a_{|\mathcal{L}|-1})}}\left(\Upsilon^{(t)}_{\mathcal{Q}(a_{|\mathcal{L}|-1})}\right)^2\right. \\ &+\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\cdots\sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})}\mathbb{1}^{(t)}_{\left\{\mathcal{Q}(a_{|\mathcal{L}|-2})\right\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})|^3\left(\lambda^{(t)}_{\mathcal{Q}(a_{|\mathcal{L}|-2})}\right)^{2\theta^{(t)}_{\mathcal{Q}(a_{|\mathcal{L}|-2})}}\left(\Upsilon^{(t)}_{\mathcal{Q}(a_{|\mathcal{L}|-2})}\right)^2+\cdots \\ &+\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\mathbb{1}^{(t)}_{\{\mathcal{Q}(a_1)\}}|\mathcal{Q}^{(k)}(a_1)|^3\left(\lambda^{(t)}_{\mathcal{Q}(a_1)}\right)^{2\theta^{(t)}_{\mathcal{Q}(a_1)}}\left(\Upsilon^{(t)}_{\mathcal{Q}(a_1)}\right)^2 \\ &\left.\left.+\mathbb{1}^{(t)}_{\{L_{1,1}\}}|\mathcal{L}^{(k)}_{1,1}|^3\left(\lambda^{(t)}_{\mathcal{L}_{1,1}}\right)^{2\theta^{(t)}_{\mathcal{L}_{1,1}}}\left(\Upsilon^{(t)}_{\mathcal{L}_{1,1}}\right)^2\Big]+\frac{4}{\eta}\left(\frac{D-D_s^{(t)}}{D}\right)^2c_1\right)\right)\end{aligned} \tag{84}$$

APPENDIX G
CLUSTER SAMPLING

In a system of a million/billion users, one technique that a main server can use to reduce the network load is to engage a portion of the devices in each global iteration. We realize this in FogL via *cluster sampling* using which at each global iteration, a portion of the clusters of the bottom-most layer are engaged in model training, which we call them as *active clusters*. We assume that at each global iteration k, the main server engages a set of $|\mathcal{S}^{(k)}|$ clusters in the learning, where each element of the set $\mathcal{S}^{(k)}$ corresponds to one cluster in the bottom-most layer. Consequently, we partition the nodes in different layers into *active nodes* (those that are through the path between an active cluster and the main server) and *passive nodes*. Similarly, for the clusters of the middle layers, if the cluster contains at least one active node, it is called an *active cluster*. To capture these dynamics, with some abuse of notation, let $\mathbb{1}^{(k)}_{\{C\}}$ take the value of 1 if cluster C is both in active mode and operates in LUT mode in global aggregation k, and 0 otherwise. To conduct analysis, in addition to our assumptions made in Assumptions 1 and 2, we also consider the following assumption that is common in stochastic optimization literature [56]:

$$\exists c_1 \geq 0, c_2 \geq 1 : \|\nabla f_i(x)\|^2 \leq c_1 + c_2\|\nabla F(x)\|^2, \quad \forall i, x. \tag{85}$$

Proposition 4. *For global iteration k of* MH-FL *with cluster sampling, the upper bound of convergence of the objective function is given by* (84)*, where $D_S^{(k)}$ denotes the total number of data points of the sampled devices at iteration k, i.e., $D_s^{(k)} = \sum_{n\in\mathcal{N}_{|\mathcal{L}|}} \mathbb{1}^{(k)}_{\{\mathcal{B}(n)\}}|\mathcal{D}_n|$, with $\mathcal{B}(n)$ referring to the cluster that node n belongs to.*[7]

Proof. To find the relationship between $\mathbf{w}^{(k)}$ and $\mathbf{w}^{(k-1)}$, we follow the procedure described in the main text. Let $\mathbb{1}^{(k)}_{\{S(C)\}}$ take the value of 1 when cluster C is in active mode in global aggregation k, and 0 otherwise. Also, with some abuse of notation, let $\mathbb{1}^{(k)}_{\{C\}}$ take the value of 1 if cluster C is both in active mode and operates in LUT mode in global aggregation k, and 0 otherwise. It can be verified that, at global iteration k, the parameter of the node located in the L_1 sampled by the main server, referred to as a_1', is given by (86), which is used by the server to obtain the next global parameter as follows:

$$\mathbf{w}^{(k)} = \frac{|\mathcal{L}^{(k)}_{1,1}|\mathbf{w}^{(k)}_{a_1'}}{D_s^{(k)}}, \tag{87}$$

where $D_s^{(k)} = \sum_{n\in\mathcal{N}_{|\mathcal{L}|}} \mathbb{1}^{(k)}_{\{\mathcal{B}(n)\}}|\mathcal{D}_n|$, with $\mathcal{B}(n)$ referring to the cluster that the node n belongs to, is the total number of data points available at the sampled devices at global aggregation k, which is assumed to be known to the server (in this case the server needs the knowledge of the number of data points available at the active clusters). Following a similar procedure described in Appendix A, we obtain (88). Let us define $\varpi^{(k)}$ as follows:

[7]It is assumed that $\prod_{j=k+1}^{k} c_j = 1$, $\forall c_j$.

$$\begin{aligned}
\widehat{\mathbf{w}}^{(k)}_{a_1'} = & \frac{\sum_{a_1\in\mathcal{L}^{(k)}_{1,1}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\mathbb{1}^{(k)}_{\{S(Q(a_{|\mathcal{L}|-1}))\}}|\mathcal{D}_{a_{|\mathcal{L}|}}|\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}}}{|\mathcal{L}^{(k)}_{1,1}|} \\
& - \frac{\sum_{a_1\in\mathcal{L}^{(k)}_{1,1}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\mathbb{1}^{(k)}_{\{S(Q(a_{|\mathcal{L}|-1}))\}}\beta|\mathcal{D}_{a_{|\mathcal{L}|}}|\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})}{|\mathcal{L}^{(k)}_{1,1}|} \\
& + \sum_{a_1\in\mathcal{L}^{(k)}_{1,1}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})}\frac{\mathbb{1}^{(k)}_{\{Q(a_{|\mathcal{L}|-1})\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})|\mathbf{c}^{(k)}_{a'_{|\mathcal{L}|}}}{|\mathcal{L}^{(k)}_{1,1}|} \\
& + \sum_{a_1\in\mathcal{L}^{(k)}_{1,1}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})}\frac{\mathbb{1}^{(k)}_{\{Q(a_{|\mathcal{L}|-2})\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})|\mathbf{c}^{(k)}_{a'_{|\mathcal{L}|-1}}}{|\mathcal{L}^{(k)}_{1,1}|} + \\
& \vdots \\
& + \sum_{a_1\in\mathcal{L}^{(k)}_{1,1}}\frac{\mathbb{1}^{(k)}_{\{Q(a_1)\}}|\mathcal{Q}^{(k)}(a_1)|\mathbf{c}^{(k)}_{a_2'}}{|\mathcal{L}^{(k)}_{1,1}|} + \mathbb{1}^{(k)}_{\{L_{1,1}\}}\mathbf{c}^{(k)}_{a_1'}
\end{aligned} \tag{86}$$

$$\begin{aligned}\mathbf{w}^{(k)} = \mathbf{w}^{(k-1)} &- \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\mathbb{1}^{(k)}_{\left\{S(Q(a_{|\mathcal{L}|-1}))\right\}}\beta\frac{|\mathcal{D}_{a_{|\mathcal{L}|}}|}{D_s^{(k)}}\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})\\
&+\frac{1}{D_s^{(k)}}\Bigg[\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})}\mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-1})\right\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})|\mathbf{c}^{(k)}_{a'_{|\mathcal{L}|}}\\
&+\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})}\mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-2})\right\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})|\mathbf{c}^{(k)}_{a'_{|\mathcal{L}|-1}}+\\
&\vdots\\
&+\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\mathbb{1}^{(k)}_{\left\{Q(a_1)\right\}}|\mathcal{Q}^{(k)}(a_1)|\mathbf{c}^{(k)}_{a'_2}+\mathbb{1}^{(k)}_{\left\{L_{1,1}\right\}}|\mathcal{L}^{(k)}_{1,1}|\mathbf{c}^{(k)}_{a'_1}\Bigg]\end{aligned}\tag{88}$$

$$\begin{aligned}\varpi^{(k)} \triangleq \frac{1}{D_s^{(k)}}&\Bigg[\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})}\mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-1})\right\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})|\mathbf{c}^{(k)}_{a'_{|\mathcal{L}|}}\\
&+\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})}\mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-2})\right\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})|\mathbf{c}^{(k)}_{a'_{|\mathcal{L}|-1}}+\\
&\vdots\\
&+\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\mathbb{1}^{(k)}_{\left\{Q(a_1)\right\}}|\mathcal{Q}^{(k)}(a_1)|\mathbf{c}^{(k)}_{a'_2}+\mathbb{1}^{(k)}_{\left\{L_{1,1}\right\}}|\mathcal{L}^{(k)}_{1,1}|\mathbf{c}^{(k)}_{a'_1}\Bigg].\end{aligned}\tag{89}$$

By adding and subtracting a term, we rewrite (88) as follows:

$$\begin{aligned}\mathbf{w}^{(k)} = \mathbf{w}^{(k-1)} &- \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\mathbb{1}^{(k)}_{\left\{S(Q(a_{|\mathcal{L}|-1}))\right\}}\beta\frac{|\mathcal{D}_{a_{|\mathcal{L}|}}|}{D_s^{(k)}}\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})\\
&+\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\beta\frac{|\mathcal{D}_{a_{|\mathcal{L}|}}|}{D}\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})\\
&-\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\beta\frac{|\mathcal{D}_{a_{|\mathcal{L}|}}|}{D}\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})+\varpi^{(k)},\end{aligned}\tag{90}$$

or equivalently

$$\begin{aligned}\mathbf{w}^{(k)} &= \mathbf{w}^{(k-1)} - \beta\nabla F(\mathbf{w}^{(k-1)})\\
&-\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\mathbb{1}^{(k)}_{\left\{S(Q(a_{|\mathcal{L}|-1}))\right\}}\beta\frac{|\mathcal{D}_{a_{|\mathcal{L}|}}|}{D_s^{(k)}}\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})\\
&+\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\beta\frac{|\mathcal{D}_{a_{|\mathcal{L}|}}|}{D}\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})+\varpi^{(k)}.\end{aligned}\tag{91}$$

Let us define $\varrho^{(k)}$ as follows:

$$\begin{aligned}\varrho^{(k)} \triangleq \beta\Bigg[&-\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\mathbb{1}^{(k)}_{\left\{S(Q(a_{|\mathcal{L}|-1}))\right\}}\frac{|\mathcal{D}_{a_{|\mathcal{L}|}}|}{D_s^{(k)}}\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})\\
&+\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\frac{|\mathcal{D}_{a_{|\mathcal{L}|}}|}{D}\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})+\frac{1}{\beta}\varpi^{(k)}\Bigg].\end{aligned}\tag{92}$$

For global iteration k, let $\bar{\mathcal{S}}^{(k)}$ denotes the set of passive clusters, which is the complementary set of $\mathcal{S}^{(k)}$, i.e., $\bar{\mathcal{S}}^{(k)}\cup\mathcal{S}^{(k)}=\mathcal{L}_{|\mathcal{L}|}$, $\bar{\mathcal{S}}^{(k)}\cap\mathcal{S}^{(k)}=\varnothing$, where $\mathcal{L}_{|\mathcal{L}|}$ denotes the set of all clusters located in the bottom-most layer. Let $\mathbb{1}^{(k)}_{\left\{\bar{S}(C)\right\}}$ take the value of 1 when cluster C is in passive mode in global aggregation k, and 0 otherwise. Following the procedure described in the proof of Appendix A, we first aim to bound $\mathbb{E}[\|\varrho^{(k)}\|^2]$. The procedure is described in (96). In that series of simplifications in (96),

$$\|\varpi^{(k)}\|^2 \leq \frac{\Phi}{\left(D_s^{(k)}\right)^2}\Bigg[\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\cdots\sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})}\mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-1})\right\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})|^3\left(\lambda^{(k)}_{Q(a_{|\mathcal{L}|-1})}\right)^{2\theta^{(k)}_{Q(a_{|\mathcal{L}|-1})}}\left(\Upsilon^{(k)}_{Q(a_{|\mathcal{L}|-1})}\right)^2 + \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\cdots\sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})}\mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-2})\right\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})|^3\left(\lambda^{(k)}_{Q(a_{|\mathcal{L}|-2})}\right)^{2\theta^{(k)}_{Q(a_{|\mathcal{L}|-2})}}\left(\Upsilon^{(k)}_{Q(a_{|\mathcal{L}|-2})}\right)^2 + \cdots + \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\mathbb{1}^{(k)}_{\{Q(a_1)\}}|\mathcal{Q}^{(k)}(a_1)|^3\left(\lambda^{(k)}_{Q(a_1)}\right)^{2\theta^{(k)}_{Q(a_1)}}\left(\Upsilon^{(k)}_{Q(a_1)}\right)^2 + \mathbb{1}^{(k)}_{\{L_{1,1}\}}|\mathcal{L}^{(k)}_{1,1}|^3\left(\lambda^{(k)}_{L_{1,1}}\right)^{2\theta^{(k)}_{L_{1,1}}}\left(\Upsilon^{(k)}_{L_{1,1}}\right)^2\Bigg] \tag{94}$$

$$F(\mathbf{w}^{(k)}) - F(\mathbf{w}^*) \leq (1-\frac{\mu}{\eta})\left(F(\mathbf{w}^{(k-1)}) - F(\mathbf{w}^*)\right) + \frac{\eta}{2}\Bigg[\frac{2\Phi}{\left(D_s^{(k)}\right)^2}\Big[\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\cdots\sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})}\mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-1})\right\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})|^3\left(\lambda^{(k)}_{Q(a_{|\mathcal{L}|-1})}\right)^{2\theta^{(k)}_{Q(a_{|\mathcal{L}|-1})}}\left(\Upsilon^{(k)}_{Q(a_{|\mathcal{L}|-1})}\right)^2 + \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\cdots\sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})}\mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-2})\right\}}|\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})|^3\left(\lambda^{(k)}_{Q(a_{|\mathcal{L}|-2})}\right)^{2\theta^{(k)}_{Q(a_{|\mathcal{L}|-2})}}\left(\Upsilon^{(k)}_{Q(a_{|\mathcal{L}|-2})}\right)^2 + \cdots + \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\mathbb{1}^{(k)}_{\{Q(a_1)\}}|\mathcal{Q}^{(k)}(a_1)|^3\left(\lambda^{(k)}_{Q(a_1)}\right)^{2\theta^{(k)}_{Q(a_1)}}\left(\Upsilon^{(k)}_{Q(a_1)}\right)^2 + \mathbb{1}^{(k)}_{\{L_{1,1}\}}|\mathcal{L}^{(k)}_{1,1}|^3\left(\lambda^{(k)}_{L_{1,1}}\right)^{2\theta^{(k)}_{L_{1,1}}}\left(\Upsilon^{(k)}_{L_{1,1}}\right)^2\Big] + \frac{8}{\eta^2}\left(\frac{D-D_s^{(k)}}{D}\right)^2(c_1+2c_2\eta(F(\mathbf{w}^{(k-1)})-F(\mathbf{w}^*))\Bigg] \tag{95}$$

the triangle inequality is applied repeatedly. In inequality (a), we have used the fact that $(\|\mathbf{a}\|+\|\mathbf{b}\|)^2 \leq 2(\|\mathbf{a}\|^2+\|\mathbf{b}\|^2)$, in inequality (b) we have used the fact that $\frac{1}{D_s^{(k)}} = \frac{1}{D} - \frac{D_s^{(k)}-D}{(D)(D_s^{(k)})}$, in (c) we have used (85), and in inequality (d) we have used the smoothness definition in Assumption 1 that can also be written as

$$F(\mathbf{y}) \leq F(\mathbf{x}) + (\mathbf{y}-\mathbf{x})^\top\nabla F(\mathbf{x}) + \frac{\eta}{2}\|\mathbf{y}-\mathbf{x}\|^2,\ \forall\mathbf{x},\mathbf{y}, \tag{93}$$

minimizing the both hand sides of which results in: $\|\nabla F(\mathbf{w})\|^2 \leq 2\eta(F(\mathbf{w})-F(\mathbf{w}^*)),\ \forall\mathbf{w}$. Note that $\|\varpi^{(k)}\|^2$ can be obtained similar to Appendix A as (94). Replacing this with $\beta=\frac{1}{\eta}$ in the bound in (96), and following the procedure of proof in Appendix A, we get (95), which can be recursively expanded to get the bound in the proposition statement. ■

Remark 2. *The methodology used to derive all the previous results regarding the convergence and the number of D2D can be studied for this scenario with cluster sampling, which we leave as future work. One key observation from* (84) *is that upon increasing the number of active clusters, often resulting in increasing* $D_s^{(k)}$, $\forall k$, *the right hand side of* (84) *starts to decrease, which implies a higher training accuracy, and the similarity between the bounds* (14) *and* (84) *increases. In the limiting case* $D_s^{(k)}=D$, $\forall k$, *bound* (84) *can be written similarly to* (84), *where* $\frac{\eta\Phi}{2D^2}$ *in* (14) *would be replaced by a larger value* $\frac{\eta\Phi}{D^2}$.

$$
\begin{aligned}
&\frac{1}{\beta^2}\|\varrho^{(k)}\|^2 = \Bigg\| - \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\mathbb{1}^{(k)}_{\{S(Q(a_{|\mathcal{L}|-1}))\}}\frac{|\mathcal{D}_{a_{|\mathcal{L}|}}|}{D_s^{(k)}}\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})\\
&+\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\frac{|\mathcal{D}_{a_{|\mathcal{L}|}}|}{D}\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})+\frac{1}{\beta}\varpi^{(k)}\Bigg\|^2\\
&\le\Bigg(\left\|\frac{1}{\beta}\varpi^{(k)}\right\|+\Bigg\|\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\frac{|\mathcal{D}_{a_{|\mathcal{L}|}}|}{D}\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})\\
&-\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\mathbb{1}^{(k)}_{\{S(Q(a_{|\mathcal{L}|-1}))\}}\frac{|\mathcal{D}_{a_{|\mathcal{L}|}}|}{D_s^{(k)}}\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})\Bigg\|\Bigg)^2\\
&\overset{(a)}{\le}2\left\|\frac{1}{\beta}\varpi^{(k)}\right\|^2+2\Bigg\|\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\frac{|\mathcal{D}_{a_{|\mathcal{L}|}}|}{D}\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})\\
&-\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\mathbb{1}^{(k)}_{\{S(Q(a_{|\mathcal{L}|-1}))\}}\frac{|\mathcal{D}_{a_{|\mathcal{L}|}}|}{D_s^{(k)}}\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})\Bigg\|^2\\
&\overset{(b)}{\le}2\frac{1}{\beta^2}\left\|\varpi^{(k)}\right\|^2+2\Bigg\|\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\mathbb{1}^{(k)}_{\{\bar{S}(Q(a_{|\mathcal{L}|-1}))\}}\frac{|\mathcal{D}_{a_{|\mathcal{L}|}}|}{D}\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})\\
&-\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\mathbb{1}^{(k)}_{\{S(Q(a_{|\mathcal{L}|-1}))\}}\frac{(D-D_s^{(k)})|\mathcal{D}_{a_{|\mathcal{L}|}}|}{(D)(D_s^{(k)})}\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})\Bigg\|^2\\
&\le2\frac{1}{\beta^2}\left\|\varpi^{(k)}\right\|^2+2\Bigg(\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\mathbb{1}^{(k)}_{\{\bar{S}(Q(a_{|\mathcal{L}|-1}))\}}\frac{|\mathcal{D}_{a_{|\mathcal{L}|}}|}{D}\left\|\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})\right\|\\
&-\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\mathbb{1}^{(k)}_{\{S(Q(a_{|\mathcal{L}|-1}))\}}\frac{(D-D_s^{(k)})|\mathcal{D}_{a_{|\mathcal{L}|}}|}{(D)(D_s^{(k)})}\left\|\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})\right\|\Bigg)^2\\
&\le2\frac{1}{\beta^2}\left\|\varpi^{(k)}\right\|^2+2\Bigg(\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\mathbb{1}^{(k)}_{\{\bar{S}(Q(a_{|\mathcal{L}|-1}))\}}\frac{|\mathcal{D}_{a_{|\mathcal{L}|}}|}{D}\max_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\left(\left\|\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})\right\|\right)\\
&+\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}}\sum_{a_2\in\mathcal{Q}^{(k)}(a_1)}\sum_{a_3\in\mathcal{Q}^{(k)}(a_2)}\cdots\sum_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\mathbb{1}^{(k)}_{\{S(Q(a_{|\mathcal{L}|-1}))\}}\frac{(D-D_s^{(k)})|\mathcal{D}_{a_{|\mathcal{L}|}}|}{(D)(D_s^{(k)})}\max_{a_{|\mathcal{L}|}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})}\left(\left\|\nabla f_{a_{|\mathcal{L}|}}(\mathbf{w}^{(k-1)}_{a_{|\mathcal{L}|}})\right\|\right)\Bigg)^2\\
&=2\frac{1}{\beta^2}\left\|\varpi^{(k)}\right\|^2+2\left(\frac{D-D_s^{(k)}}{D}\max_{a\in\mathcal{N}_{|\mathcal{L}|}}\left(\left\|\nabla f_a(\mathbf{w}_a^{(k-1)})\right\|\right)+\frac{(D-D_s^{(k)})D_s^{(k)}}{(D)(D_s^{(k)})}\max_{a\in\mathcal{N}_{|\mathcal{L}|}}\left(\left\|\nabla f_a(\mathbf{w}_a^{(k-1)})\right\|\right)\right)^2\\
&\le\frac{2}{\beta^2}\left\|\varpi^{(k)}\right\|^2+2\left[\left(2\frac{D-D_s^{(k)}}{D}\max_{a\in\mathcal{N}_{|\mathcal{L}|}}\left(\left\|\nabla f_a(\mathbf{w}_a^{(k-1)})\right\|\right)\right)^2\right]\le\frac{2}{\beta^2}\left\|\varpi^{(k)}\right\|^2+8\left(\frac{D-D_s^{(k)}}{D}\max_{a\in\mathcal{N}_{|\mathcal{L}|}}\left(\left\|\nabla f_a(\mathbf{w}_a^{(k-1)})\right\|\right)\right)^2\\
&\le\frac{2}{\beta^2}\left\|\varpi^{(k)}\right\|^2+8\left(\frac{D-D_s^{(k)}}{D}\right)^2\left(\max_{a\in\mathcal{N}_{|\mathcal{L}|}}\left(\left\|\nabla f_a(\mathbf{w}_a^{(k-1)})\right\|\right)\right)^2\\
&\overset{(c)}{\le}\frac{2}{\beta^2}\left\|\varpi^{(k)}\right\|^2+8\left(\frac{D-D_s^{(k)}}{D}\right)^2\left(c_1+c_2\|F(\mathbf{w}^{(k-1)})\|^2\right)\\
&\overset{(d)}{\le}2\frac{1}{\beta^2}\left\|\varpi^{(k)}\right\|^2+8\left(\frac{D-D_s^{(k)}}{D}\right)^2\left(c_1+2c_2\eta\left(F(\mathbf{w}^{(k-1)})-F(\mathbf{w}^*)\right)\right)
\end{aligned}
\tag{96}
$$

The following appendix is the last appendix of the paper concerned with theoretical analysis, which is followed by another appendix containing extensive numerical simulations.

APPENDIX H
AGGREGATION ERROR UPON USING ALGORITHM 3

According to (32), the aggregation error at the k-th global aggregation is given by

$$\begin{aligned}\mathbf{e}^{(k)} = \frac{1}{D}\Bigg(& \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \sum_{a_2\in\mathcal{Q}^{(k)}(a_1)} \sum_{a_3\in\mathcal{Q}^{(k)}(a_2)} \cdots \sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})} \mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-1})\right\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})| \mathbf{c}^{(k)}_{a'_{|\mathcal{L}|}} \\ & + \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \sum_{a_2\in\mathcal{Q}^{(k)}(a_1)} \sum_{a_3\in\mathcal{Q}^{(k)}(a_2)} \cdots \sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})} \mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-2})\right\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})| \mathbf{c}^{(k)}_{a'_{|\mathcal{L}|-1}} + \\ & \vdots \\ & + \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \mathbb{1}^{(k)}_{\{Q(a_1)\}} |\mathcal{Q}^{(k)}(a_1)| \mathbf{c}^{(k)}_{a'_2} + \mathbb{1}^{(k)}_{\{L_{1,1}\}} |\mathcal{L}^{(k)}_{1,1}| \mathbf{c}^{(k)}_{a'_1} \Bigg).\end{aligned} \tag{97}$$

Following a similar procedure described in Appendix A, we get

$$\begin{aligned}\|\mathbf{e}^{(k)}\|^2 \leq \frac{\Phi}{D^2} \Bigg[& \\ & \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \sum_{a_2\in\mathcal{Q}^{(k)}(a_1)} \cdots \sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})} \mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-1})\right\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-1})|^3 \left(\lambda^{(k)}_{Q(a_{|\mathcal{L}|-1})}\right)^{2\theta^{(k)}_{Q(a_{|\mathcal{L}|-1})}} \left(\Upsilon^{(k)}_{Q(a_{|\mathcal{L}|-1})}\right)^2 \\ & + \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \sum_{a_2\in\mathcal{Q}^{(k)}(a_1)} \cdots \sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})} \mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-2})\right\}} |\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})|^3 \left(\lambda^{(k)}_{Q(a_{|\mathcal{L}|-2})}\right)^{2\theta^{(k)}_{Q(a_{|\mathcal{L}|-2})}} \left(\Upsilon^{(k)}_{Q(a_{|\mathcal{L}|-2})}\right)^2 \\ & + \cdots + \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \mathbb{1}^{(k)}_{\{Q(a_1)\}} |\mathcal{Q}^{(k)}(a_1)|^3 \left(\lambda^{(k)}_{Q(a_1)}\right)^{2\theta^{(k)}_{Q(a_1)}} \left(\Upsilon^{(k)}_{Q(a_1)}\right)^2 + \mathbb{1}^{(k)}_{\{L_{1,1}\}} |\mathcal{L}^{(k)}_{1,1}|^3 \left(\lambda^{(k)}_{L_{1,1}}\right)^{2\theta^{(k)}_{L_{1,1}}} \left(\Upsilon^{(k)}_{L_{1,1}}\right)^2 \Bigg],\end{aligned} \tag{98}$$

where $\Phi = N_{|\mathcal{L}|-1} + N_{|\mathcal{L}|-2} + \cdots + N_1 + 1$. By tuning the number of D2D according to (30), following a similar procedure as Appendix B, we get

$$\begin{aligned}\|\mathbf{e}^{(k)}\|^2 & \leq \frac{\Phi}{D^2} \Bigg[\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \sum_{a_2\in\mathcal{Q}^{(k)}(a_1)} \cdots \sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})} \mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-1})\right\}} \frac{\psi}{\frac{\Phi}{D^2} N_{|\mathcal{L}|-1}|\mathcal{L}|} \\ & + \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \sum_{a_2\in\mathcal{Q}^{(k)}(a_1)} \cdots \sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})} \mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-2})\right\}} \frac{\psi}{\frac{\Phi}{D^2} N_{|\mathcal{L}|-2}|\mathcal{L}|} + \cdots \\ & + \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \mathbb{1}^{(k)}_{\{Q(a_1)\}} \frac{\psi}{\frac{\Phi}{D^2} N_1|\mathcal{L}|} + \mathbb{1}^{(k)}_{\{L_{1,1}\}} \frac{\psi}{\frac{\Phi}{D^2} N_0|\mathcal{L}|} \Bigg] \\ & \leq \frac{\Phi}{D^2} \Bigg[\sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \sum_{a_2\in\mathcal{Q}^{(k)}(a_1)} \cdots \sum_{a_{|\mathcal{L}|-1}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-2})} \frac{\psi}{\frac{\Phi}{D^2} N_{|\mathcal{L}|-1}|\mathcal{L}|} \\ & + \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \sum_{a_2\in\mathcal{Q}^{(k)}(a_1)} \cdots \sum_{a_{|\mathcal{L}|-2}\in\mathcal{Q}^{(k)}(a_{|\mathcal{L}|-3})} \mathbb{1}^{(k)}_{\left\{Q(a_{|\mathcal{L}|-2})\right\}} \frac{\psi}{\frac{\Phi}{D^2} N_{|\mathcal{L}|-2}|\mathcal{L}|} + \cdots \\ & + \sum_{a_1\in\mathcal{L}_{1,1}^{(k)}} \frac{\psi}{\frac{\Phi}{D^2} N_1|\mathcal{L}|} + \frac{\psi}{\frac{\Phi}{D^2} N_0|\mathcal{L}|} \Bigg] \\ & = \frac{\Phi}{D^2} \Bigg[\underbrace{\frac{\psi}{\frac{\Phi}{D^2}|\mathcal{L}|} + \frac{\psi}{\frac{\Phi}{D^2}|\mathcal{L}|} \cdots + \frac{\psi}{\frac{\Phi}{D^2}|\mathcal{L}|} + \frac{\psi}{\frac{\Phi}{D^2}|\mathcal{L}|}}_{|\mathcal{L}| \text{ terms}} \Bigg].\end{aligned} \tag{99}$$

Thus, we have

$$\|\mathbf{e}^{(k)}\|^2 \leq \psi. \tag{100}$$

APPENDIX I
DETAILS OF THE SIMULATIONS SETTING AND FURTHER SIMULATIONS

In this section, we first present some details regarding simulations settings and parameter tuning and then present a series of simulation results regarding the choice of different datasets and larger network size as compared to the main text. Our entire Python implementation, including the set of hyperparameters used in each experiment, can be found at the following Github repository: https://github.com/shams-sam/Federated2Fog".

A. Simulation Setting

1) Setup: All simulations are performed on a single machine with 64GB RAM and 8GB GPU memory, which emulates the learning through a distributed learning framework *PySyft* that helps spin off virtually disjoint nodes with mutually exclusive model parameters and datasets, working on top of *PyTorch* machine learning library.

2) Classifiers: We consider two different classifiers - regularized Support Vector Machine (SVM) and fully-connected Neural Network (NN), initialized with a copy of global model before the learning process begins on each node participating in the learning process.

The regularized SVM is tuned to satisfy the strong convexity with $\mu = 0.1$. We also use the estimated value of $\eta = 10$ (similar values are observed in [19]). The NN classifier is a simple fully connected network with a single hidden layer and no convolutional units. *Softmax* activation at the output layer gives the class *logits* and the overall training optimizes negative log-likelihood loss function with L2 regularization.

Input size for both the models, SVM and NN is $28 \times 28 = 784$, with output size 10. The number of parameters optimized by the networks M is given by $M = (784 + 1) \times 10 = 7850$.

3) Datasets and Data Distribution among the Nodes: We consider two datasets MNIST and F-MNIST (Fashion MNIST)[8], each of which contain 60000 training samples and 10000 testing samples. MNIST consists of handwritten digits $0 - 9$, while F-MNIST consists of images associated with 10 classes in clothing domain. Each dataset consists of 28×28 grayscale images.

The datasets are distributed over nodes such that all nodes have approximately equal number of training samples. However the training samples, maybe either be i.i.d or non-i.i.d distributed. For i.i.d distribution, each node participating in the learning process has samples from each class of the dataset, while under non-i.i.d distribution, each node has access to only one of the classes. These are the extreme ends of possible split of data among nodes in terms of class distribution, helping us evaluate the overall robustness as well as differences in characteristics of our technique under different settings.

4) Network Formation: We consider two network configurations: (i) the network consists of 125 edge devices; (ii) the network consists of 625 edge devices. For the former case, we consider a fog network consisting of a main server and three sub-layers, to build our fog network we start with the 125 worker nodes in the bottom-most layer (L_3) and dedicated local datasets sampled as explained above. The worker nodes update the local models with a copy of parameters from latest global model at the start of each iteration. The worker nodes are then clustered in groups of 5 to communicate with one of the 25 aggregators in their upper layer (i.e., L_2), such that there is a 1-to-1 mapping between the clusters and the aggregators. Similarly the nodes in layer (L_2) are clustered and communicate with the 5 aggregators in the layer L_1, followed by clustering and communicating the 5 nodes with the main server.

For the latter case, we consider a fog network consisting of a main server and four sub-layers, to build our fog network we start with the 625 worker nodes in the bottom-most layer (i.e., L_4) and dedicated local datasets sampled as explained above. The worker nodes update the local models with a copy of parameters from latest global model at the start of each iteration. The worker nodes are then clustered in groups of 5 to communicate with one of the 125 aggregators in their upper layer (i.e., L_3), such that there is a 1-to-1 mapping between the clusters and aggregators. Similarly nodes in L_3 are again clustered and communicate to the 25 aggregators in the upper layer (i.e., L_2). This is followed by clustering of these nodes in groups of 5 and communicating with 5 aggregators in layer L_1, which then communicate with the main server.

The connectivity among the nodes within a cluster is simulated using random geometric graphs with increasing connectivity as we traverse from the bottom-most layer to the main server. In our random geometric graph construction, nodes are placed in a circle disc with radius 100m uniformly at random, where the existence of edge (i.e., D2D link) between two nodes is assumed if the distance between the nodes is less than a threshold (φ). For the case with 125 edge device, layer L_3 has $\varphi = 40$m, followed by layer L_2 with $\varphi = 50$m and layer L_1 with $\varphi = 60$m. For the case with 625 edge device, in layer L_3 and L_4 we have $\varphi = 40$, followed by layer L_2 with $\varphi = 50$ and layer L_1 with $\varphi = 60$. We use *NetworkX*[9] library of *Python* for generating the graph. We adjust the radius parameter of the graph generator such that the average degree of the graph is within tolerance region of 0.2 from the desired degree of the graph.

For the D2D communications, we consider the common choice of the weights [46] that gives $\mathbf{z}_n^{(t+1)} = \mathbf{z}_n^{(t)} + d_C^{(k)} \sum_{m \in \zeta^{(k)}(n)} (\mathbf{z}_m^{(t)} - \mathbf{z}_n^{(t)})$, $0 < d_C^{(k)} < 1/D_C^{(k)}$, for any node n in arbitrary cluster C, where $D_C^{(k)}$ is the maximum

[8]https://github.com/zalandoresearch/fashion-mnist

[9]https://networkx.github.io

degree of the nodes in $G_C^{(k)}$. Using this implementation, the nodes inside LUT cluster C only need to have the knowledge of the parameter $d_C^{(k)}$, which is broadcast by the respective parent node.

We summarize the simulation parameters in Table I.

Table I: Summary of parameter values employed in our simulations.

Parameter	Value
Number of Edge Devices	125, 625
Number of Layers of the Network	4, 5
Number of Devices Per Cluster	5
Random Geometric Graph Threshold φ	$[40, 60]m$
Smoothness η	10
Strong Convexity μ	0.1
Number of Data points D	$60,000$
Uplink Transmit Power of Devices	24dBm
D2D Transmit Power of Devices	10dBm
D2D/Uplink Delay of Transmission of Parameters	0.25 Sec

B. Further Simulation Results

This section presents the plots from complimentary experiments from Section IV. In the following, we explain the relationship between the figures presented in this appendix and the simulation results presented in the main text.

Fig. 5 from main text is repeated in Fig. 16 for MNIST dataset distributed over 625 edge devices, Fig. 27 for FMNIST dataset distributed over 125 edge devices and Fig. 38 for FMNIST dataset distributed over 625 edge devices.

Fig. 6 from main text is repeated in Fig. 17 for MNIST dataset distributed over 625 edge devices, Fig. 28 for FMNIST dataset distributed over 125 edge devices and Fig. 39 for FMNIST dataset distributed over 625 edge devices.

Fig. 7 from main text is repeated in Fig. 18 for MNIST dataset distributed over 625 edge devices, Fig. 29 for FMNIST dataset distributed over 125 edge devices and Fig. 40 for FMNIST dataset distributed over 625 edge devices.

Fig. 8 from main text is repeated in Fig. 19 for MNIST dataset distributed over 625 edge devices, Fig. 30 for FMNIST dataset distributed over 125 edge devices and Fig. 41 for FMNIST dataset distributed over 625 edge devices.

Fig. 9 from main text is repeated in Fig. 20 for MNIST dataset distributed over 625 edge devices, Fig. 31 for FMNIST dataset distributed over 125 edge devices and Fig. 42 for FMNIST dataset distributed over 625 edge devices.

Fig. 10 from main text is repeated in Fig. 21 for MNIST dataset distributed over 625 edge devices, Fig. 32 for FMNIST dataset distributed over 125 edge devices and Fig. 43 for FMNIST dataset distributed over 625 edge devices.

Fig. 11 from main text is repeated in Fig. 22 for MNIST dataset distributed over 625 edge devices, Fig. 33 for FMNIST dataset distributed over 125 edge devices and Fig. 44 for FMNIST dataset distributed over 625 edge devices.

Fig. 12 from main text is repeated in Fig. 23 for MNIST dataset distributed over 625 edge devices, Fig. 34 for FMNIST dataset distributed over 125 edge devices and Fig. 45 for FMNIST dataset distributed over 625 edge devices.

Fig. 13 from main text is repeated in Fig. 24 for MNIST dataset distributed over 625 edge devices, Fig. 35 for FMNIST dataset distributed over 125 edge devices and Fig. 46 for FMNIST dataset distributed over 625 edge devices.

Fig. 14 from main text is repeated in Fig. 25 for MNIST dataset distributed over 625 edge devices, Fig. 36 for FMNIST dataset distributed over 125 edge devices and Fig. 47 for FMNIST dataset distributed over 625 edge devices.

Fig. 15 from main text is repeated in Fig. 26 for MNIST dataset distributed over 625 edge devices, Fig. 37 for FMNIST dataset distributed over 125 edge devices and Fig. 48 for FMNIST dataset distributed over 625 edge devices.

C. Energy and Parameter Transmission Savings under Various D2D Control Parameters and Tolerable Aggregation Errors

Varying σ: To demonstrate the effect of σ on the energy and data traffic savings, we set σ_j at layer L_j as $\sigma_j = \sigma' \max_i \Upsilon^{(1)}_{L_{j,i}}$, where $\Upsilon^{(1)}_{L_{j,i}}$ is the divergence of parameters at the beginning of model training at i-th cluster of layer j, and change the value of σ' in our experiments. Note that higher values of $\{\sigma_j\}_{j=1}^{|\mathcal{L}|}$ are associated with a looser condition on the D2D consensus formation error (see Proposition 1 and 2). This implies that increasing $\{\sigma_j\}_{j=1}^{|\mathcal{L}|}$ often results in performing fewer D2D rounds across the fog layers. The results for varying values of σ' are depicted in Fig. 49 and 50 (for MNIST dataset and 125 nodes); Fig. 51 and 52 (for MNIST dataset and 625 nodes); Fig. 53 and 54 (for FMNIST dataset and 125 nodes); Fig. 55 and 56 (for FMNIST dataset and 625 nodes).

- Considering the energy consumption (i.e., Figs. 49,51,53,55), increasing σ' often results in more energy savings since the nodes conduct less D2D rounds. However, after a certain threshold, increasing σ' may lead to slight increase in energy consumption for `MH-FL` (e.g., increasing σ' from 0.6 to 0.9 in Fig. 49). That is because decreasing the D2D

communication rounds below a threshold may have a significantly negative impact on the convergence speed, where the model may need considerably higher number of global aggregation iterations to reach the desired accuracy.

- Considering the parameter transmission savings (i.e., Figs. 50,52,54,56), it can be noted that increasing σ' often results in a slight increase in parameter transmission for MH−FL, since the model may need a few more global aggregations to reach the desired accuracy when the D2D rounds conducted are decreased. Note that in all the cases, MH−FL outperforms the EUT baseline method in terms of both energy consumption and parameter transmissions.

Varying ψ: Note that parameter ψ controls the 2-norm of aggregation errors when MH−FL is used with non-convex loss functions. In particular, smaller values of ψ impose a smaller tolerable error of aggregation at the server, which call for higher number of D2D rounds across the nodes to decrease the local aggregation errors. The results are depicted in Fig. 57 and 58 (for MNIST dataset and 125 nodes); Fig. 59 and 60 (for MNIST dataset and 625 nodes); Fig. 61 and 62 (for FMNIST dataset and 125 nodes); Fig. 63 and 64 (for FMNIST dataset and 625 nodes).

- Considering the energy consumption (i.e., Figs. 57,59,61,63), increasing ψ results in more energy savings since the nodes conduct fewer D2D rounds. Also, it can be seen that for small values of ψ (e.g. $\psi = 10$ in these figures), MH−FL has a higher energy consumption as compared to EUT baseline since the number of D2D communication rounds become unreasonably high for such choices of ψ. However for moderate to high value of ψ (e.g., $\psi \geq 10^3$ in these figures), MH−FL always outperforms the EUT baseline in terms of energy consumption.
- Considering the parameter transmission savings (i.e., Figs. 58,60,62,64), increasing ψ often results in increasing the number of parameter transmissions for MH−FL since the model may need more time (i.e., higher number of global aggregations) to reach the desired accuracy. Nevertheless, MH−FL outperforms the EUT baseline in all the scenarios due the sampling of a single node from each LUT cluster.

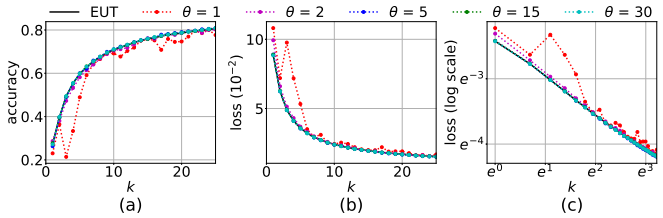


Fig. 16: Performance comparison between baseline EUT and MH−MT when a fixed number of D2D rounds θ is used at every cluster of the network, for non-i.i.d. As the number of D2D rounds increases, MH−MT performs more similar to the EUT baseline and the learning is more stable. (MNIST, 625 Edge Devices)

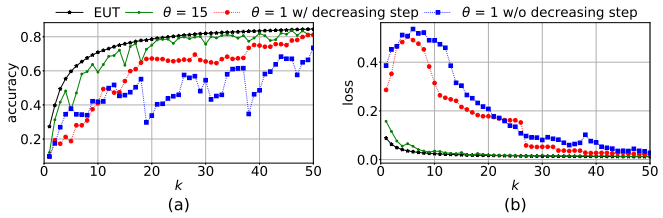


Fig. 17: Performance comparison between baseline EUT, and MH−MT with and without (w/o) decreasing the gradient descent step size. Decreasing the step size can provide convergence to the optimal solution in cases where a fixed step size is not capable, but also has a slower convergence speed. (MNIST, 625 Edge Devices)

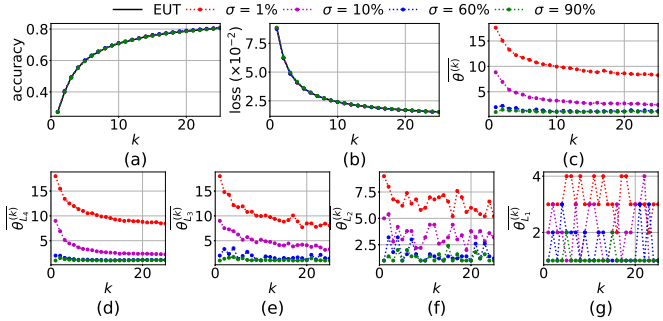


Fig. 18: Performance comparison between baseline EUT and MH−MT for i.i.d when a finite optimality gap is tolerable. σ_j at L_j is fixed as $\sigma_j = \sigma' \max_i \Upsilon^{(1)}_{L_{j,i}}$. Tapering of D2D rounds through time and space (layers) can be observed. (MNIST, 625 Edge Devices)

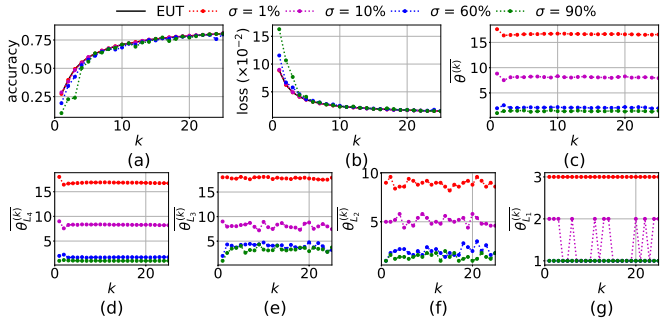


Fig. 19: Performance comparison between baseline EUT and MH−MT for non-i.i.d when a finite optimality gap is tolerable. σ_i is set as in Fig. 18. Smaller loss and higher accuracy are achieved with smaller σ', implying more rounds of consensus. (MNIST, 625 Edge Devices)

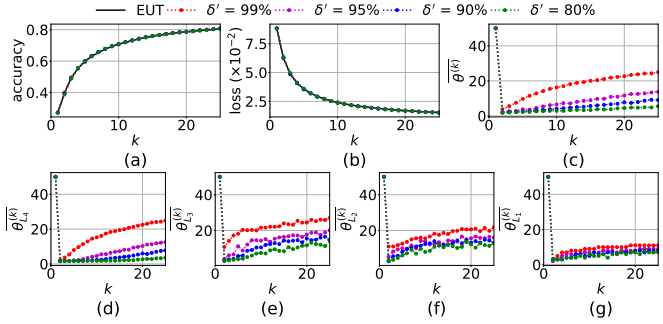


Fig. 20: Performance comparison between baseline EUT and MH−MT for i.i.d. when linear convergence to the optimal is desired. The value of δ is set at $\delta = \delta' \frac{\mu}{\eta}$. Boosting of the D2D rounds through time can be observed. Also, tapering through space can be observed by comparing the D2D rounds in the bottom subplots. (MNIST, 625 Edge Devices)

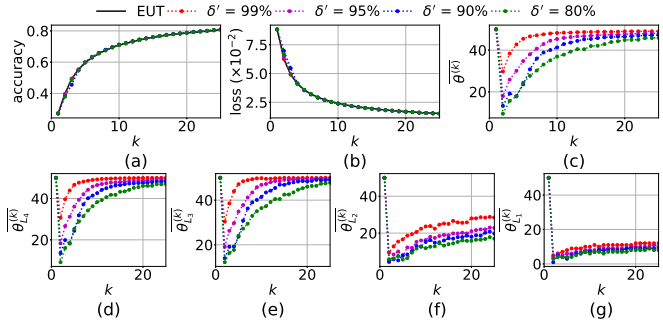


Fig. 21: Performance comparison between baseline EUT and MH−MT for non-i.i.d when linear convergence to the optimal is desired. The value of δ is set as in Fig. 20. Smaller values of loss and higher accuracy are both associated with larger value of δ, which results in lower error tolerance and more rounds of consensus. (MNIST, 625 Edge Devices)

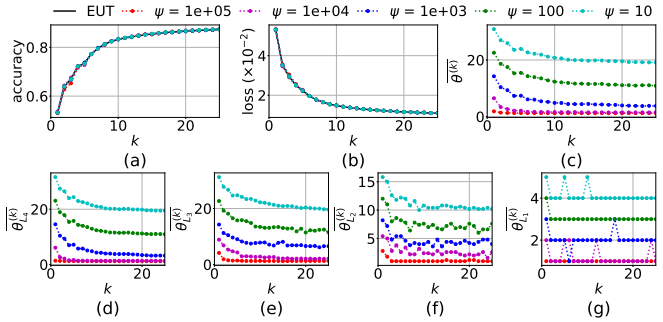


Fig. 22: Performance comparison between baseline EUT and MH−MT under i.i.d using NNs with different values of ψ. Tapering the D2D rounds through time can be observed. Also, tapering through space can be observed by comparing the D2D rounds in the bottom subplots. (MNIST, 625 Edge Devices)

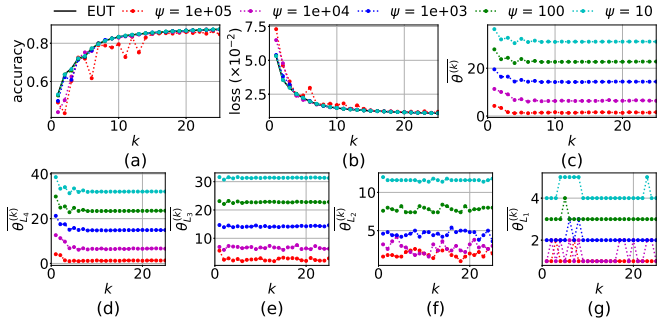


Fig. 23: Performance comparison between baseline EUT and MH−MT under non-i.i.d. using NNs with different values of ψ. Lower loss and higher accuracy are associated with smaller values of ψ, which result in lower error tolerance and larger values of D2D rounds over time. (MNIST, 625 Edge Devices)

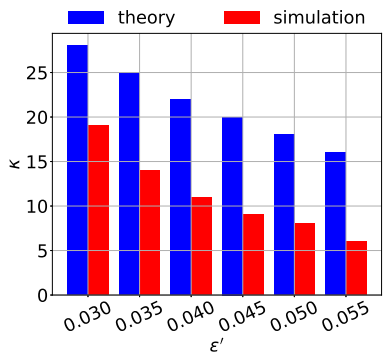


Fig. 24: Comparison between the theoretical and simulation results regarding the number of global iterations to achieve an accuracy of $\epsilon'(F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*))$ for different ϵ'. Convergence in practice is faster than the derived upper bound. (MNIST, 625 Edge Devices)

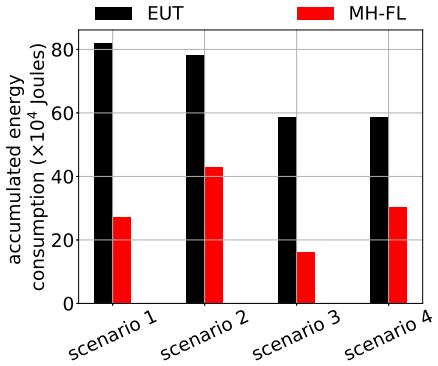


Fig. 25: Comparison of accumulated energy consumption between EUT and MH−MT over scenario 1: $\sigma' = 0.1$ from Fig. 18, scenario 2: $\sigma' = 0.1$ from Fig. 19, scenario 3: $\psi = 10^4$ from Fig. 22, and scenario 4: $\psi = 10^4$ from Fig. 23. (MNIST, 625 Edge Devices)

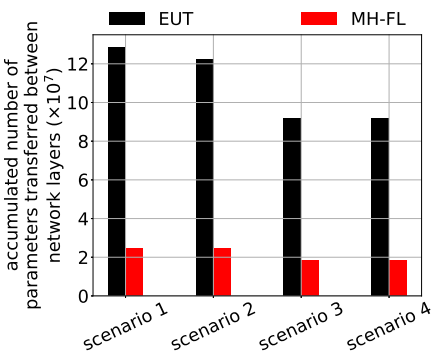


Fig. 26: Comparison of parameters transferred among layers in EUT vs MH−MT over scenario 1: $\sigma' = 0.1$ from Fig. 18, scenario 2: $\sigma' = 0.1$ from Fig. 19, scenario 3: $\psi = 10^4$ from Fig. 22, and scenario 4: $\psi = 10^4$ from Fig. 23. (MNIST, 625 Edge Devices)

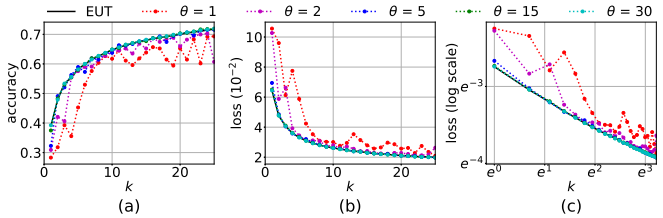


Fig. 27: Performance comparison between baseline EUT and MH−MT when a fixed number of D2D rounds θ is used at every cluster of the network, for non-i.i.d. As the number of D2D rounds increases, MH−MT performs more similar to the EUT baseline and the learning is more stable. (FMNIST, 125 Edge Devices)

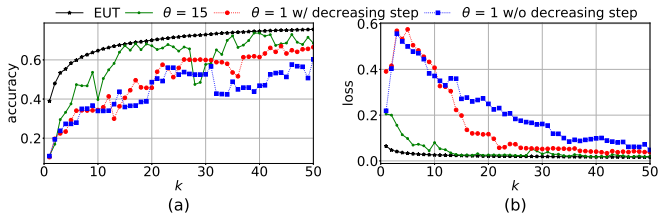


Fig. 28: Performance comparison between baseline EUT, and MH−MT with and without (w/o) decreasing the gradient descent step size. Decreasing the step size can provide convergence to the optimal solution in cases where a fixed step size is not capable, but also has a slower convergence speed. (FMNIST, 125 Edge Devices)

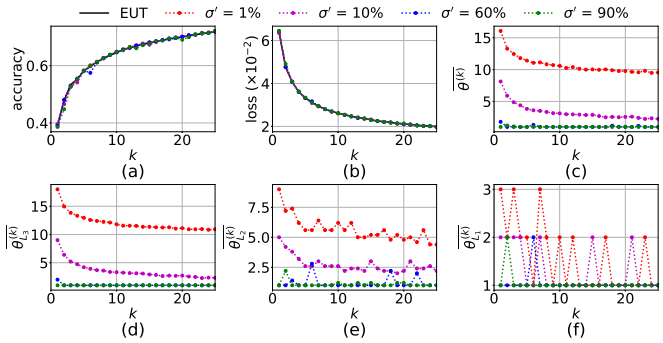


Fig. 29: Performance comparison between baseline EUT and MH−MT for i.i.d when a finite optimality gap is tolerable. σ_j at L_j is fixed as $\sigma_j = \sigma' \max_i \Upsilon^{(1)}_{L_{j,i}}$. Tapering of D2D rounds through time and space (layers) can be observed. (FMNIST, 125 Edge Devices)

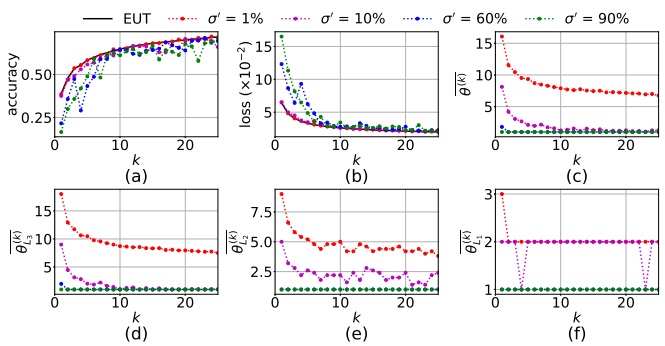


Fig. 30: Performance comparison between baseline EUT and MH−MT for non-i.i.d when a finite optimality gap is tolerable. σ_i is set as in Fig. 29. Smaller loss and higher accuracy are achieved with smaller σ', implying more rounds of consensus. (FMNIST, 125 Edge Devices)

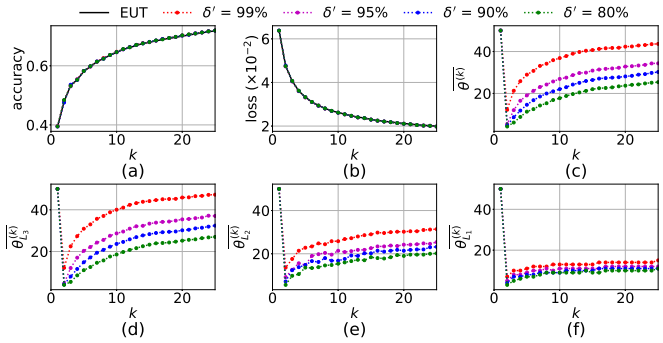


Fig. 31: Performance comparison between baseline EUT and MH−MT for i.i.d. when linear convergence to the optimal is desired. The value of δ is set at $\delta = \delta' \frac{\mu}{\eta}$. Boosting of the D2D rounds through time can be observed. Also, tapering through space can be observed by comparing the D2D rounds in the bottom subplots. (FMNIST, 125 Edge Devices)

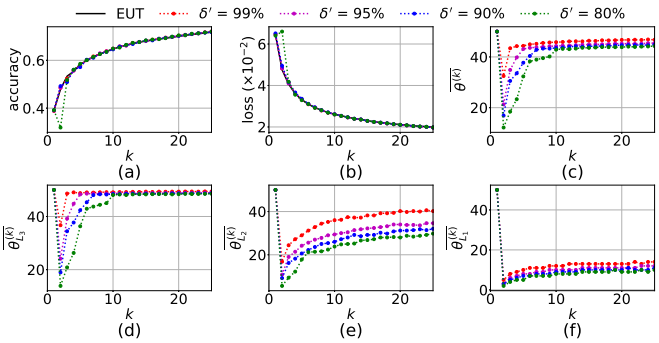


Fig. 32: Performance comparison between baseline EUT and MH−MT for non-i.i.d when linear convergence to the optimal is desired. The value of δ is set as in Fig. 31. Smaller values of loss and higher accuracy are both associated with larger value of δ, which results in lower error tolerance and more rounds of consensus. (FMNIST, 125 Edge Devices)

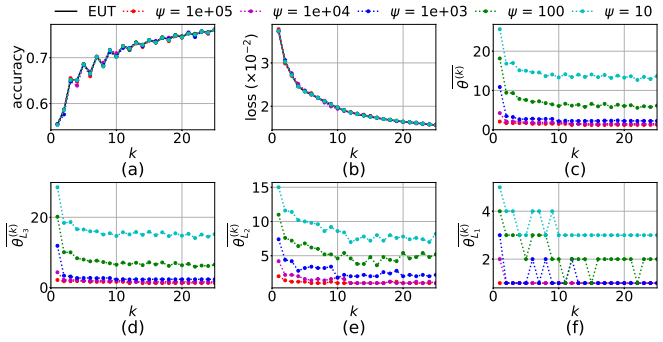


Fig. 33: Performance comparison between baseline EUT and MH−MT under i.i.d using NNs with different values of ψ. Tapering the D2D rounds through time can be observed. Also, tapering through space can be observed by comparing the D2D rounds in the bottom subplots. (FMNIST, 125 Edge Devices)

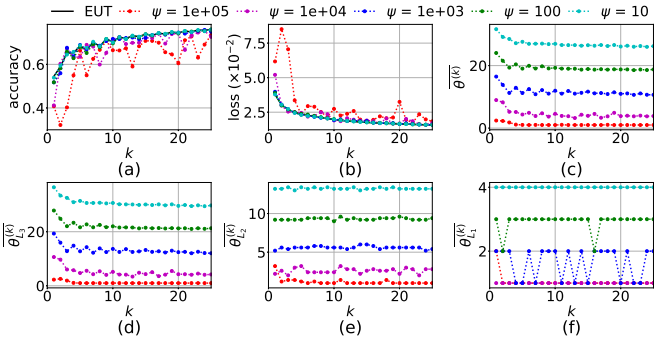


Fig. 34: Performance comparison between baseline EUT and MH−MT under non-i.i.d. using NNs with different values of ψ. Lower loss and higher accuracy are associated with smaller values of ψ, which result in lower error tolerance and larger values of D2D rounds over time. (FMNIST, 125 Edge Devices)

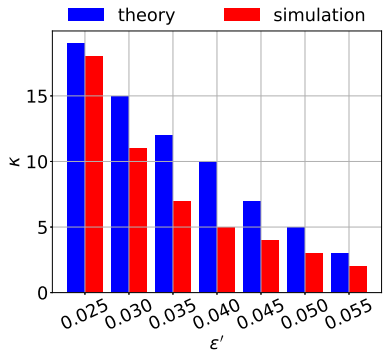


Fig. 35: Comparison between the theoretical and simulation results regarding the number of global iterations to achieve an accuracy of $\epsilon'(F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*))$ for different ϵ'. Convergence in practice is faster than the derived upper bound. (FMNIST, 125 Edge Devices)

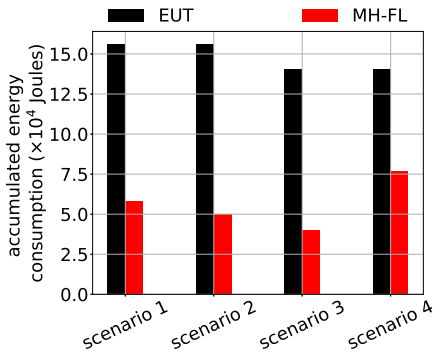


Fig. 36: Comparison of accumulated energy consumption between EUT and MH−MT over scenario 1: $\sigma' = 0.1$ from Fig. 29, scenario 2: $\sigma' = 0.1$ from Fig. 30, scenario 3: $\psi = 10^4$ from Fig. 33, and scenario 4: $\psi = 10^4$ from Fig. 34. (FMNIST, 125 Edge Devices)

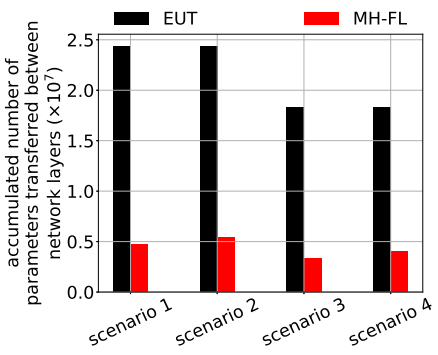


Fig. 37: Comparison of parameters transferred among layers in EUT vs MH−MT over scenario 1: $\sigma' = 0.1$ from Fig. 29, scenario 2: $\sigma' = 0.1$ from Fig. 30, scenario 3: $\psi = 10^4$ from Fig. 33, and scenario 4: $\psi = 10^4$ from Fig. 34. (FMNIST, 125 Edge Devices)

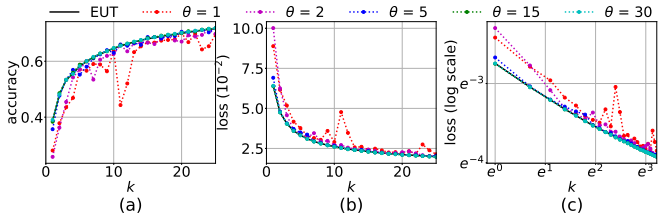


Fig. 38: Performance comparison between baseline EUT and MH−MT when a fixed number of D2D rounds θ is used at every cluster of the network, for non-i.i.d. As the number of D2D rounds increases, MH−MT performs more similar to the EUT baseline and the learning is more stable. (FMNIST, 625 Edge Devices)

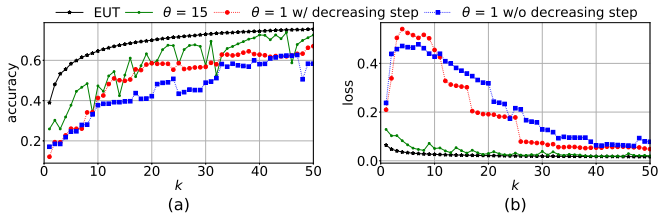


Fig. 39: Performance comparison between baseline EUT, and MH−MT with and without (w/o) decreasing the gradient descent step size. Decreasing the step size can provide convergence to the optimal solution in cases where a fixed step size is not capable, but also has a slower convergence speed. (FMNIST, 625 Edge Devices)

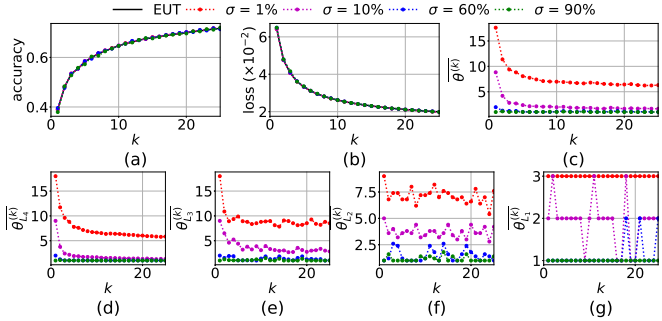


Fig. 40: Performance comparison between baseline EUT and MH−MT for i.i.d when a finite optimality gap is tolerable. σ_j at L_j is fixed as $\sigma_j = \sigma' \max_i \Upsilon^{(1)}_{L_{j,i}}$. Tapering of D2D rounds through time and space (layers) can be observed. (FMNIST, 625 Edge Devices)

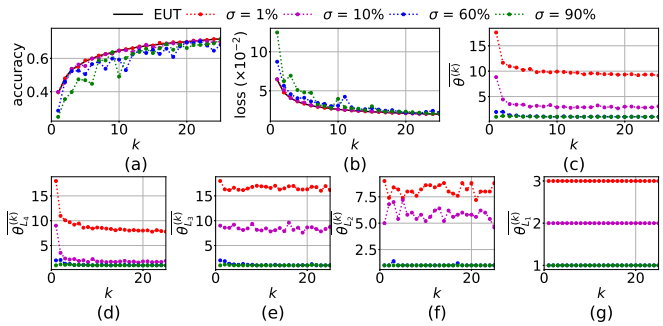


Fig. 41: Performance comparison between baseline EUT and MH−MT for non-i.i.d when a finite optimality gap is tolerable. σ_i is set as in Fig. 40. Smaller loss and higher accuracy are achieved with smaller σ', implying more rounds of consensus. (FMNIST, 625 Edge Devices)

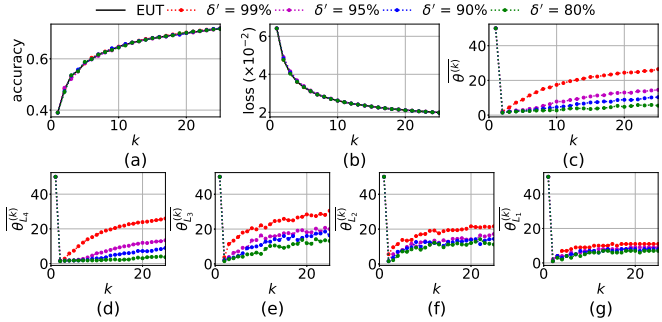


Fig. 42: Performance comparison between baseline EUT and MH−MT for i.i.d. when linear convergence to the optimal is desired. The value of δ is set at $\delta = \delta' \frac{\mu}{\eta}$. Boosting of the D2D rounds through time can be observed. Also, tapering through space can be observed by comparing the D2D rounds in the bottom subplots. (FMNIST, 625 Edge Devices)

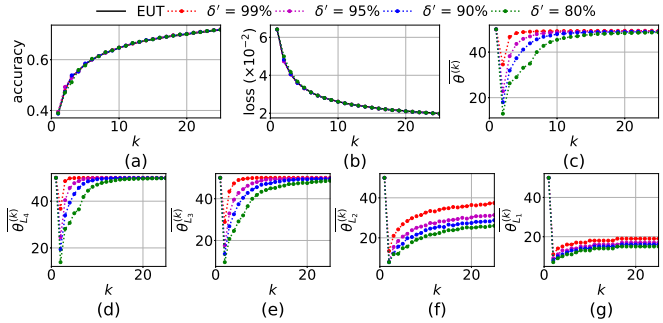


Fig. 43: Performance comparison between baseline EUT and MH−MT for non-i.i.d when linear convergence to the optimal is desired. The value of δ is set as in Fig. 42. Smaller values of loss and higher accuracy are both associated with larger value of δ, which results in lower error tolerance and more rounds of consensus. (FMNIST, 625 Edge Devices)

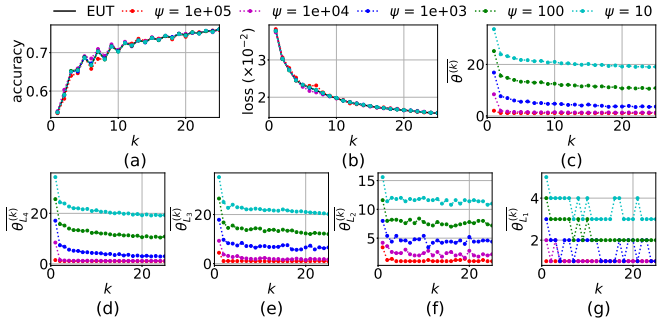


Fig. 44: Performance comparison between baseline EUT and MH−MT under i.i.d using NNs with different values of ψ. Tapering the D2D rounds through time can be observed. Also, tapering through space can be observed by comparing the D2D rounds in the bottom subplots. (FMNIST, 625 Edge Devices)

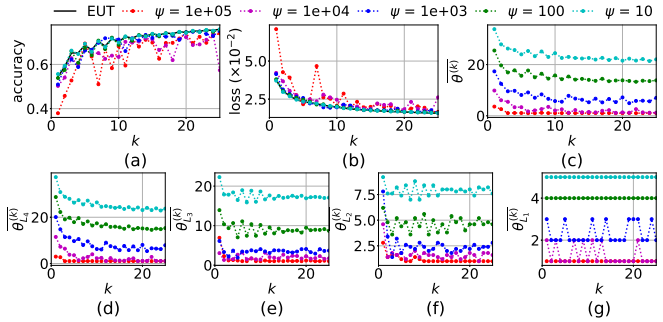


Fig. 45: Performance comparison between baseline EUT and MH−MT under non-i.i.d. using NNs with different values of ψ. Lower loss and higher accuracy are associated with smaller values of ψ, which result in lower error tolerance and larger values of D2D rounds over time. (FMNIST, 625 Edge Devices)

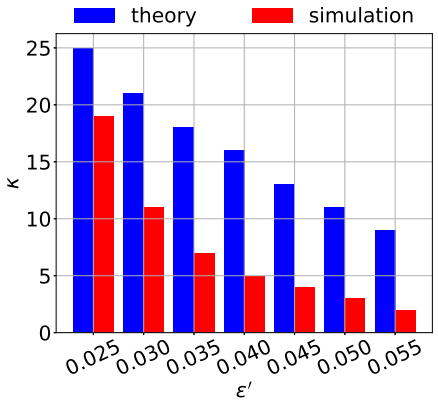


Fig. 46: Comparison between the theoretical and simulation results regarding the number of global iterations to achieve an accuracy of $\epsilon'(F(\mathbf{w}^{(0)}) - F(\mathbf{w}^*))$ for different ϵ'. Convergence in practice is faster than the derived upper bound. (FMNIST, 625 Edge Devices)

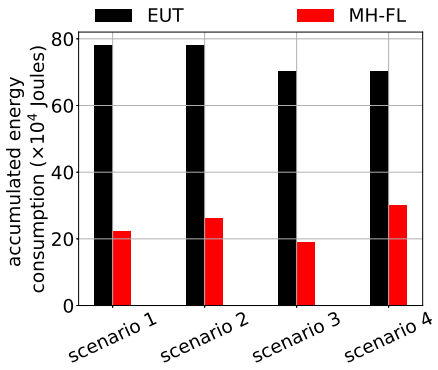


Fig. 47: Comparison of accumulated energy consumption between EUT and MH-MT over scenario 1: $\sigma' = 0.1$ from Fig. 40, scenario 2: $\sigma' = 0.1$ from Fig. 41, scenario 3: $\psi = 10^4$ from Fig. 44, and scenario 4: $\psi = 10^4$ from Fig. 45. (FMNIST, 625 Edge Devices)

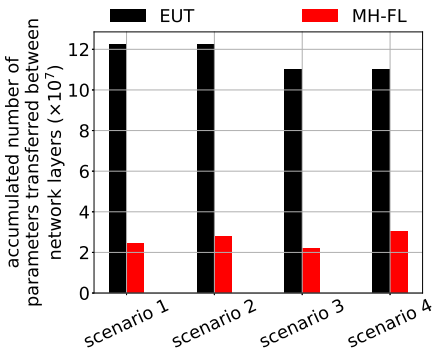


Fig. 48: Comparison of parameters transferred among layers in EUT vs MH-MT over scenario 1: $\sigma' = 0.1$ from Fig. 40, scenario 2: $\sigma' = 0.1$ from Fig. 41, scenario 3: $\psi = 10^4$ from Fig. 44, and scenario 4: $\psi = 10^4$ from Fig. 45. (FMNIST, 625 Edge Devices)

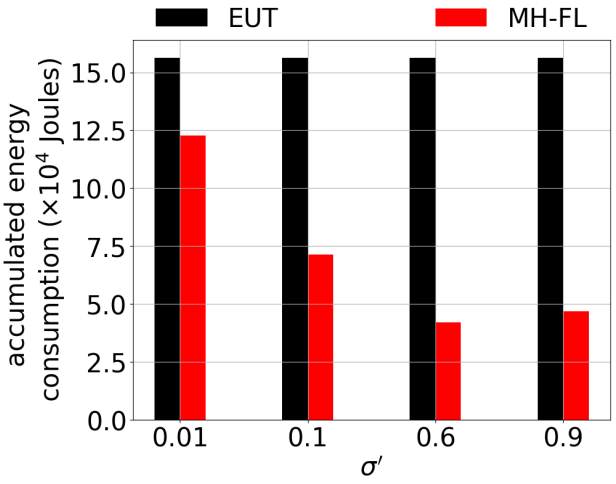


Fig. 49: Energy consumption on MNIST w/ 125 nodes for varying σ'.

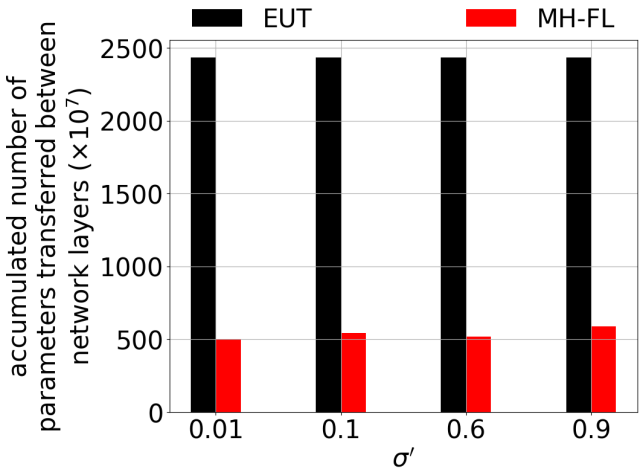


Fig. 50: Parameters transferred on MNIST w/ 125 nodes for varying σ'.

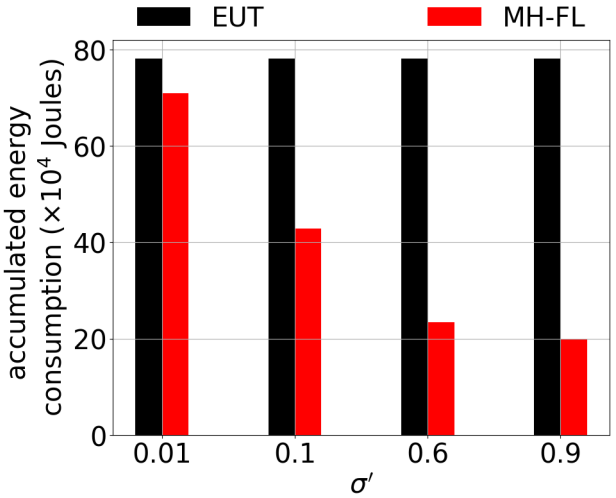


Fig. 51: Energy consumption on MNIST w/ 625 nodes for varying σ'.

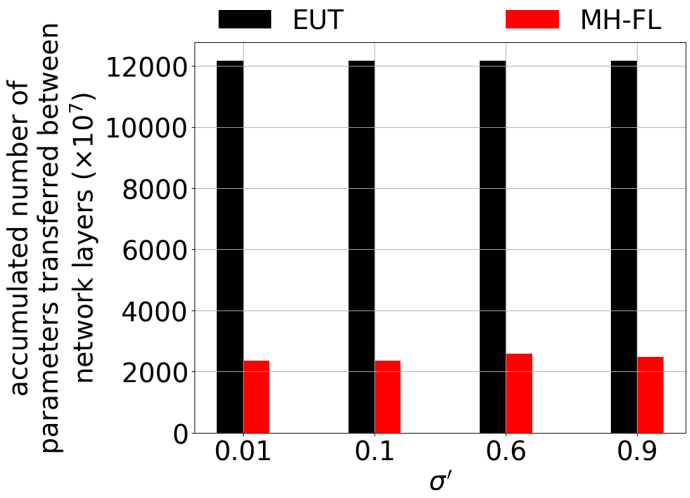


Fig. 52: Parameters transferred on MNIST w/ 625 nodes for varying σ'.

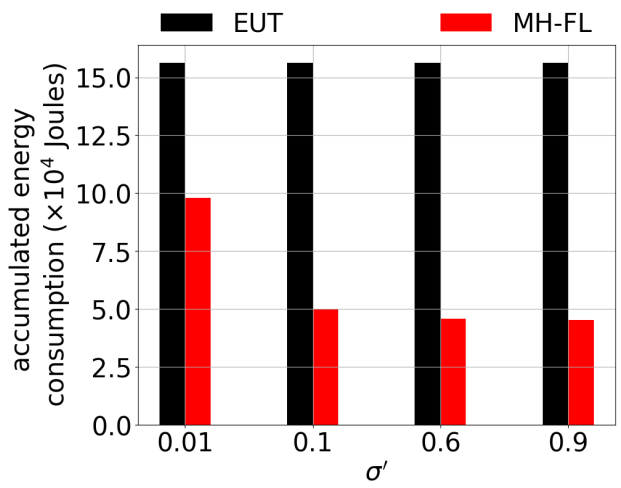


Fig. 53: Energy consumption on FMNIST w/ 125 nodes for varying σ'.

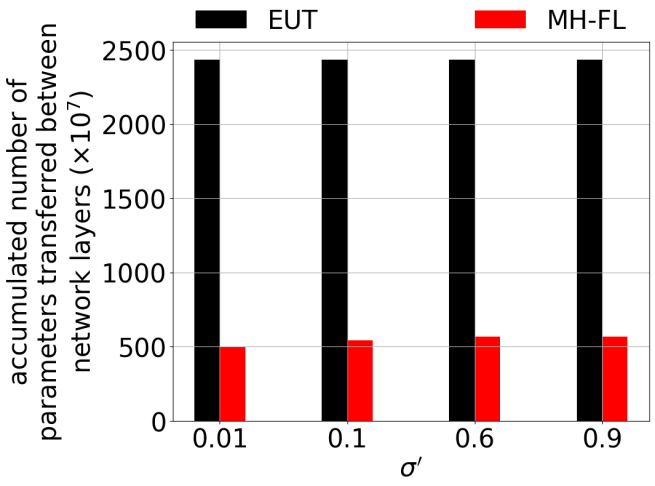


Fig. 54: Parameters transferred on FMNIST w/ 125 nodes for varying σ'.

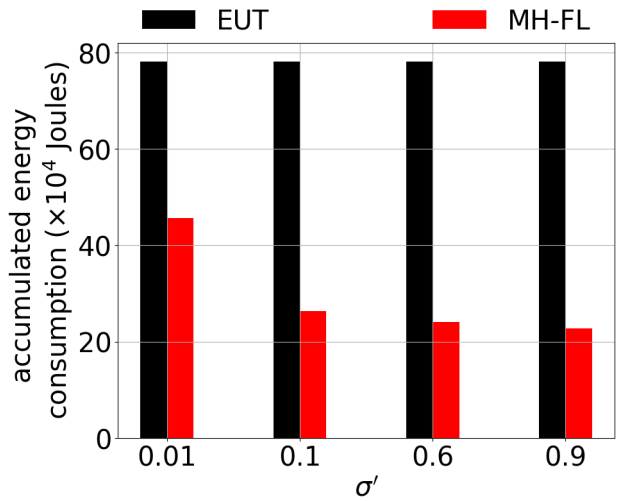


Fig. 55: Energy consumption on FMNIST w/ 625 nodes for varying σ'.

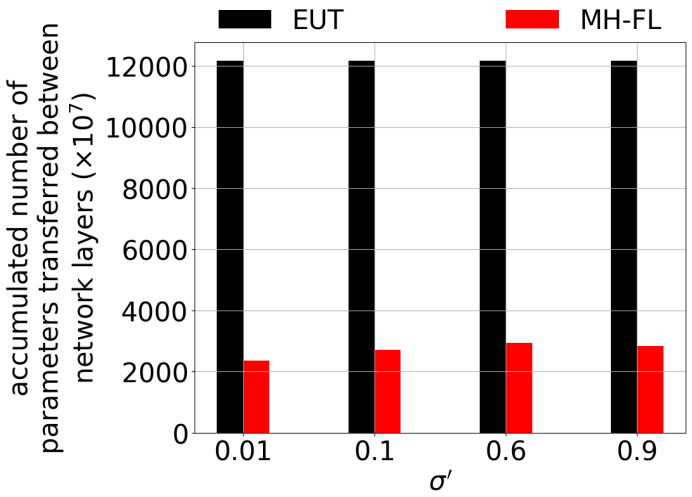


Fig. 56: Parameters transferred on FMNIST w/ 625 nodes for varying σ'.

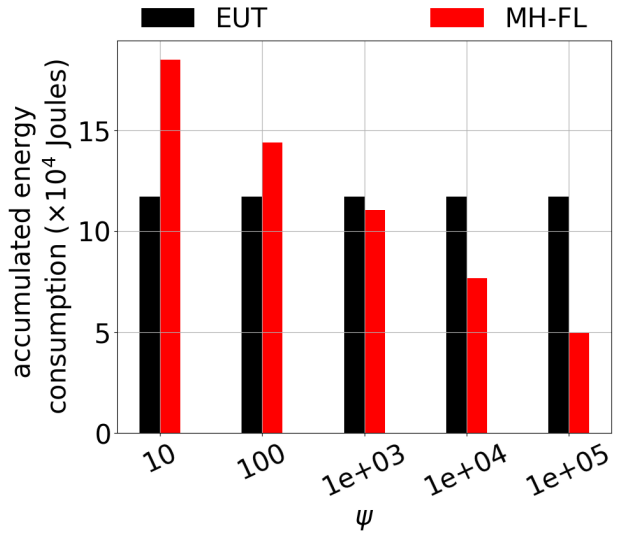


Fig. 57: Energy consumption on MNIST w/ 125 nodes for varying ψ.

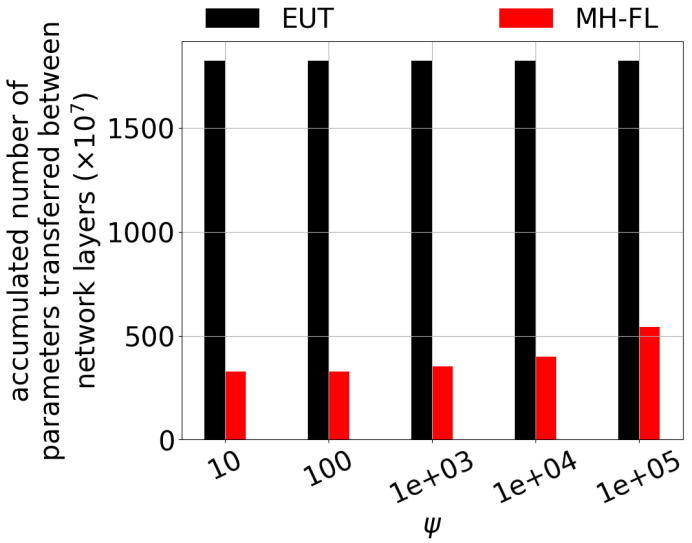


Fig. 58: Parameters transferred on MNIST w/ 125 nodes for varying ψ.

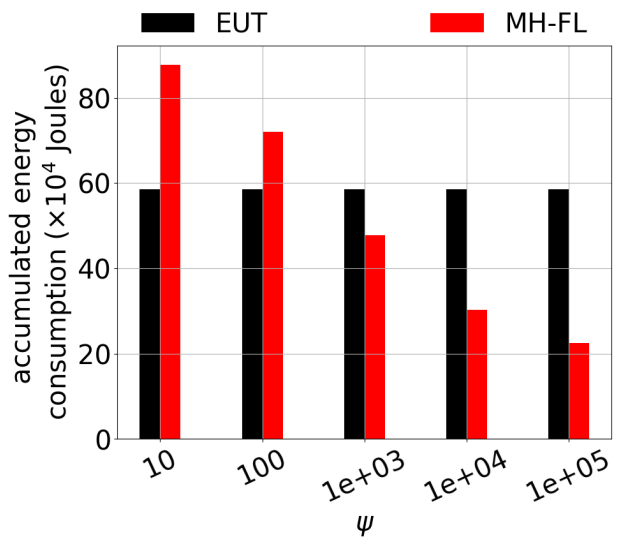


Fig. 59: Energy consumption on MNIST w/ 625 nodes for varying ψ.

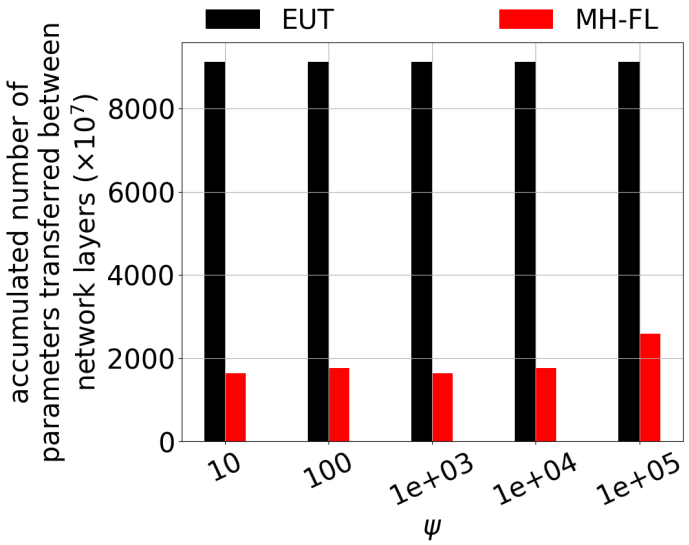


Fig. 60: Parameters transferred on MNIST w/ 625 nodes for varying ψ.

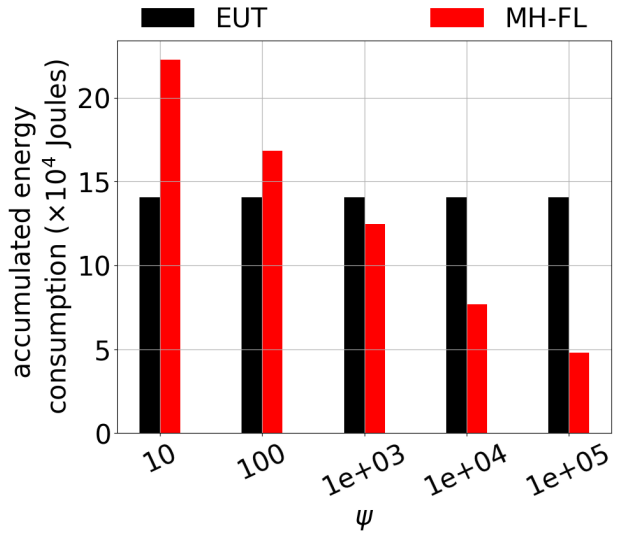


Fig. 61: Energy consumption on FMNIST w/ 125 nodes for varying ψ.

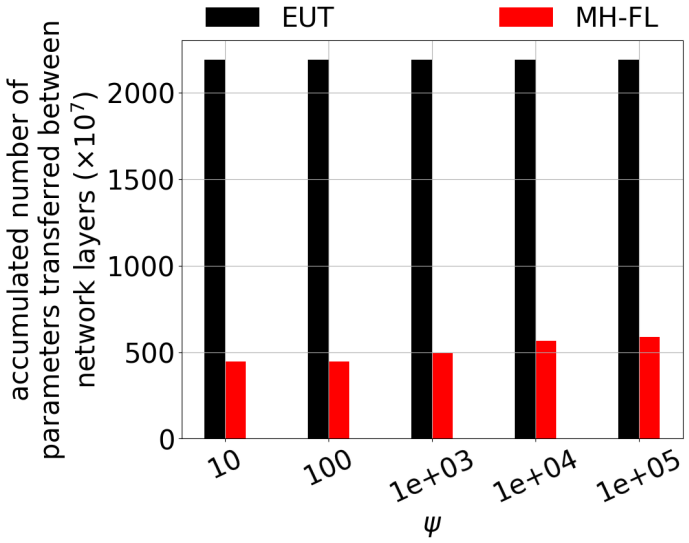


Fig. 62: Parameters transferred on FMNIST w/ 125 nodes for varying ψ.

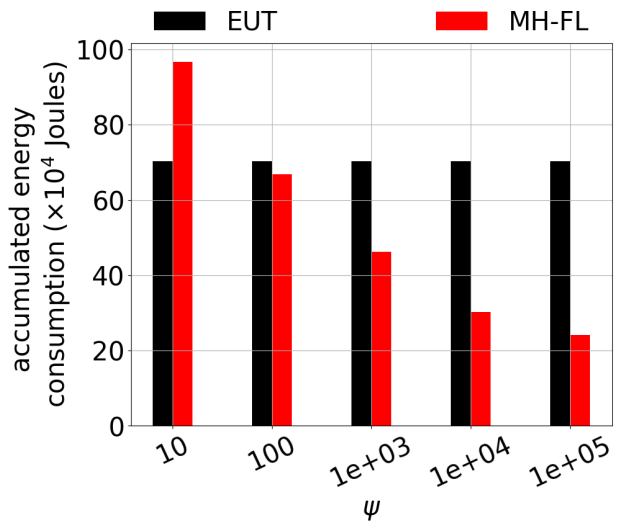


Fig. 63: Energy consumption on FMNIST w/ 625 nodes for varying ψ.

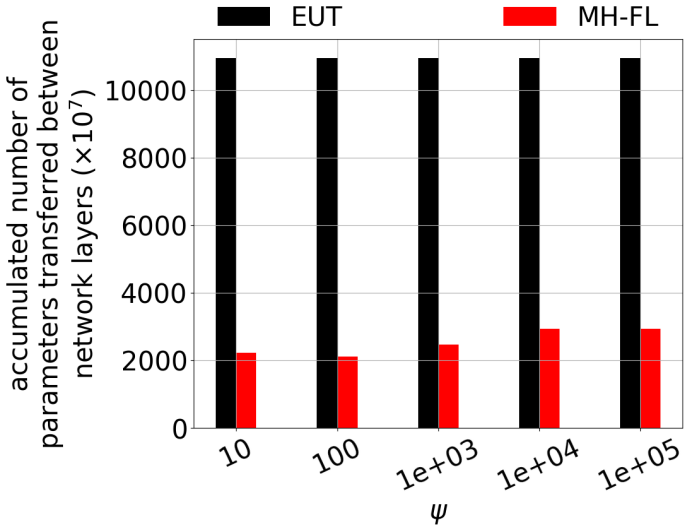


Fig. 64: Parameters transferred on FMNIST w/ 625 nodes for varying ψ.