Decentralized Federated Learning: Balancing Communication and Computing Costs

Wei Liu, Li Chen, and Wenyi Zhang, *Senior Member, IEEE*

Abstract—Decentralized stochastic gradient descent (SGD) is a driving engine for decentralized federated learning (DFL). The performance of decentralized SGD is jointly influenced by inter-node communications and local updates. In this paper, we propose a general DFL framework, which implements both multiple local updates and multiple inter-node communications periodically, to strike a balance between communication efficiency and model consensus. It can provide a general decentralized SGD analytical framework. We establish strong convergence guarantees for the proposed DFL algorithm without the assumption of convex objectives. The convergence rate of DFL can be optimized to achieve the balance of communication and computing costs under constrained resources. For improving communication efficiency of DFL, compressed communication is further introduced to the proposed DFL as a new scheme, named DFL with compressed communication (C-DFL). The proposed C-DFL exhibits linear convergence for strongly convex objectives. Experiment results based on MNIST and CIFAR-10 datasets illustrate the superiority of DFL over traditional decentralized SGD methods and show that C-DFL further enhances communication efficiency.

Index Terms—Compressed communication, decentralized machine learning, federated learning.

I. INTRODUCTION

Recently, thanks to the explosive growth of the powerful individual computing devices worldwide and the rapid advancement of Internet of Things (IoT), data generated at device terminals are experiencing an exponential increase [1]. Data-driven machine learning is developing into an attractive technique, which makes full use of the tremendous data for predictions and decisions of future events. As a promising data-driven machine learning variant, Federated Learning (FL) provides a communication-efficient approach for processing voluminous distributed data and grows in popularity nowadays.

As a critical technique of distributed machine learning, FL was first proposed in [2] under a centralized form, where edge clients perform local model training in parallel and a central server aggregates the trained model parameters from the edge without transmission of raw data from edge clients to the central server. For convergence analysis of FL, the work in [3] provided a theoretical convergence bound of FL based on gradient-descent. The area of FL has been studied in the literature from both theoretical and applied perspectives [4–6]. Because of the property of transmitting model parameters instead of user data and the distributed network structure that an arbitrary number of edge nodes are coordinated through one central server, FL plays a key role in supporting privacy protection of user data and deploying in complex environment with massive intelligent terminals access to the network center.

In order to improve learning performance with guaranteeing communication efficiency of FL, extended methods for executing FL have been proposed and developed. As a fundamental means of achieving FL, parallel SGD has been studied and developed in previous works [7] [8]. In original parallel SGD [9], each round consists of one local update (computing) step and one global aggregation (communication) step. To improve communication efficiency of distributed learning, Federated Averaging (FedAvg), as a variant of parallel SGD, was proposed in [2]. Multiple local updates are executed on each client between each communication step in FedAvg. FedAvg was generalized to non-i.i.d. cases [10–12]; momentum FL was proposed for convergence acceleration [13]; FedAvg with a Proximal Term (FedProx) was studied for handling heterogeneous federated learning environment [14]; federated multi-task learning was proposed to improve learning performance under unbalance data [15]; robust FL studied the noise effect during model transmission in wireless communication [16].

The aforementioned centralized FL framework requires a central server for data aggregation. As an aggregation center of information from clients, the server could potentially incur a single point of failure, such as suffering from high bandwidth and device battery costs [17]. It may also become the busiest node incurring a high communication cost when a large number of clients are deployed in a federated learning system [18]. When there is no central server, the DFL paradigm is rapidly gaining popularity, where edge clients exchange their local models or gradient information with their neighboring clients to achieve model consensus. This avoids a single point of failure, and reduces the communication traffic jam due to the central server [19]. DFL can be traced back to decentralized optimization [20–27] and decentralized SGD [18, 28–32] where gossip-based model averaging was applied for model consensus evaluation [33]. For decentralized optimization, subgradient descent method [20], decentralized consensus optimization with asynchrony and delays [21], alternating direction method of multipliers (ADMM) [22] [23], dual averaging method [24], Byzantine-resilient distributed coordinate descent [25] and Newton tracking method [26] were studied to solve the decentralized optimization problem, and convex optimization problem with a sum of smooth convex functions and a non-smooth regularization term was solved in [27]. As a driving engine, decentralized SGD provides a means of realizing DFL. A convergence analysis of decentralized gradient descent was provided in [28]; the convergence of decentralized approach with delayed gradient information was analyzed [29]; a comparison between decentralized SGD and centralized ones was made, illustrating the advantage of

decentralized approaches [18]. Decentralized SGD performing one local update step and one inter-node communication step in a round (D-SGD) was proposed and analyzed in [30]. In order to improve communication efficiency of decentralized SGD, the work in [31] first proposed a framework named Cooperative SGD (C-SGD), where the edge clients perform multiple local updates before exchanging the model parameters. The subsequent work in [32] studied how to efficiently combine local updates and decentralized SGD based on the decentralized framework by performing several local updates and several decentralized SGDs.

Note that decentralized SGD in the existing literature mainly focuses on improving communication efficiency by performing multiple local update steps and one inter-node communication step in a round [31]. Besides communication efficiency, model consensus also has an important influence on the convergence performance of DFL. Model consensus, which denotes the removal of discrepancy of model parameter averagings among neighboring nodes, improves the convergence performance of DFL by executing multiple inter-node communications in the DFL framework. However, to the best of our knowledge, a joint consideration of communication efficiency and model consensus in DFL framework has not been addressed. Only considering communication efficiency incurs inferior model consensus, and vice versa. Motivated by the above observation, we consider a unified DFL framework where each node performs *multiple* local updates and *multiple* inter-node communications in a round alternately. The proposed framework can balance communication efficiency and model consensus with convergence guarantee. We prove the global convergence and derive a convergence bound of the proposed DFL, which indicates that multiple inter-node communications effectively improve the convergence performance. In order to improve communication efficiency of DFL, we further introduce communication compression into DFL, and the DFL framework with communication compression is termed C-DFL. The convergence rate of C-DFL is derived and linear convergence is exhibited for strongly convex objectives. The main contributions of this paper are summarized as follows:

- **Design of DFL:** To consider the joint influence of communication efficiency and model consensus, we design a novel DFL framework to balance the allocation of communication and computation resources for optimizing the convergence performance. In DFL, each node performs multiple local update steps and multiple inter-node communication steps in a round.
- **Convergence Analysis of DFL:** We establish the convergence of DFL and derive a convergence bound that incorporates network topology and the allocation of computation and communication steps in a round. It shows that the convergence performance of DFL outperforms that of traditional frameworks.
- **C-DFL Design and Convergence:** In order to improve communication efficiency of DFL, we further introduce compressed communication to DFL as a new scheme, termed C-DFL, where inter-node communication is achieved with compressed information to reduce the
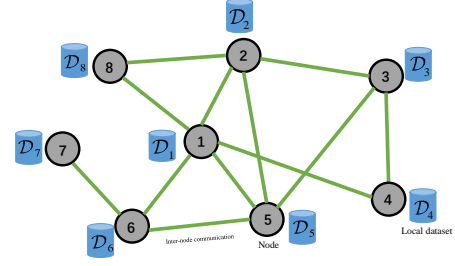

Fig. 1: System model of decentralized federated learning

communication overhead. For C-DFL, the convergence performance is based on the synergistic influence of compression and allocation of communication and computation steps in a round. We provide a convergence analysis of C-DFL and establish the linear convergence property.

The remaining part of this paper is organized as follows. We introduce the system model for solving the distributed optimization problem in decentralized scenarios in Section II. The design of DFL is described in detail in Section III and the convergence analysis of DFL is presented in Section IV. Then we introduce C-DFL and present its convergence analysis in Section V. We present the simulation results in Section VI. The conclusion is given in Section VII.

A. Notations

In this paper, we use $\mathbf{1}$ to denote vector $[1, 1, ..., 1]^\top$, and define the consensus matrix $\mathbf{J} \triangleq \mathbf{1}\mathbf{1}^\top/(\mathbf{1}^\top\mathbf{1})$, which means under the topology represented by $\mathbf{J}$, DFL can realize model consensus. All vectors in this paper are column vectors. We assume that the decentralized network contains N nodes. So we have $\mathbf{1} \in \mathbb{R}^N$ and $\mathbf{J} \in \mathbb{R}^{N \times N}$. We use $\|\cdot\|$, $\|\cdot\|_\mathrm{F}$, $\|\cdot\|_\mathrm{op}$ and $\|\cdot\|_2$ to denote the ℓ_2 vector norm, the Frobenius matrix norm, the operator norm and ℓ_2 matrix norm, respectively.

II. System Model

Considering the case of a general system model as shown in Fig. 1, we first illustrate the decentralized federated learning structure. The system model is made of N edge nodes. The N nodes have the distributed datasets $\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_i, ..., \mathcal{D}_N$ with $\mathcal{D}_i$ owned by node i. We use $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_N\}$ to denote the global dataset of all nodes. Assuming $\mathcal{D}_i \cap \mathcal{D}_j = \emptyset$ for $i \neq j$, we define $D \triangleq |\mathcal{D}|$ and $D_i \triangleq |\mathcal{D}_i|$ for $i = 1, 2, ..., N$, where $|\cdot|$ denotes the size of set, and use $\mathbf{w} \in \mathbb{R}^d$ to denote the model parameter.

Then we introduce the loss function of the decentralized network as follows. Let $(\mathbf{x}_j, y_j)$ denote a data sample, where $\mathbf{x}_j$ is regarded as a input of the learning model and y_j is the label of the input $\mathbf{x}_j$, which denotes the desired output of the model. The same learning models are embedded in all different nodes. We use $f(\cdot)$ to denote the loss function based on one data sample. Furthermore, we present the definition of loss

function based on local dataset $\mathcal{D}_i$ and the global dataset $\mathcal{D}$, which are denoted by $F_i(\cdot)$ and $F(\cdot)$, respectively. We define

$$F_i(\mathbf{w}) \triangleq \frac{1}{D_i} \sum_{j \in D_i} f(\mathbf{w}, \mathbf{x}_j, y_j),$$

$$F(\mathbf{w}) \triangleq \frac{1}{D} \sum_{i=1}^{N} D_i F_i(\mathbf{w}).$$

In order to learn about the characteristic information hidden in the distributed datasets, the purpose of DFL is to obtain optimal model parameters of all nodes. We present the learning problem of DFL. The targeted optimization problem is the minimization of the global loss function to get the optimal parameter $\mathbf{w}^*$ as follows:

$$\mathbf{w}^* = \underset{\mathbf{w} \in \mathbb{R}^d}{\arg\min}\, F(\mathbf{w}). \tag{1}$$

Because of the intrinsic complexity of most machine learning models and common training datasets, finding a closed-form solution of the optimization problem (1) is usually intractable. Thus, gradient-based method is generally used to solve this problem. We let $\mathbf{w}_t$ denote the model parameter at the t-th iteration, η denote the learning rate, $\xi_t^{(i)} \subset \mathcal{D}_i$ denote the mini-batches randomly sampled on $\mathcal{D}_i$, and

$$g(\mathbf{w}, \xi^{(i)}) \triangleq \frac{1}{|\xi^{(i)}|} \sum_{(\mathbf{x}_j, y_j) \in \xi^{(i)}} \nabla f(\mathbf{w}, \mathbf{x}_j, y_j)$$

be the stochastic gradient of node i. The iteration of the gradient-based method can be expressed as

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \left[\frac{1}{N} \sum_{i=1}^{N} g(\mathbf{w}_t, \xi_t^{(i)}) \right]. \tag{2}$$

We will use $g(\mathbf{w})$ to substitute $g(\mathbf{w}, \xi^{(i)})$ in the remaining part of the paper for convenience.

Finally, the targeted optimization problem can be collaboratively solved by local updates and model averagings. In local update step, each node uses gradient-based method to update its local model. In model averaging step, each node communicates its local model parameters with its neighboring nodes in parallel.

III. Design of DFL

In this section, we introduce the design of DFL to solve the optimization problem as presented in (1). Note that there is no central server in the system model of DFL as shown in Fig. 1. We first elaborate the motivation of our work. Then we present the design of DFL in detail. Finally, we introduce D-SGD and C-SGD as the special cases of DFL.

A. Motivation

Considering the system model, D-SGD [18] [30], where each edge node performs one local update and one inter-node communication alternately in a round, was proposed to solve the problem in (1). It will incur a lower communication efficiency. To improve communication efficiency of the model, C-SGD [31] was proposed, where each node performs multiple
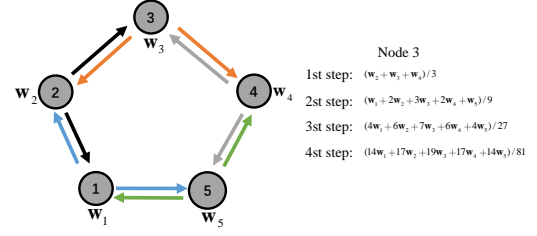

Fig. 2: Illustration of DFL with 5 nodes. Focusing on Node 3, we can find its model parameter change with different communication steps.
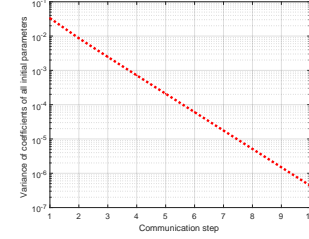

Fig. 3: Illustration of how the variance of coefficients of all initial parameters decreases monotonically with communication step at Node 3 in Fig. 2.

local updates and one inter-node communication. However, in C-SGD, communication efficiency is merely considered by performing multiple local updates in a round. It can incur local drift and inferior model consensus because successive gradient updates in a round make the descent direction locally optimal (not globally optimal). Furthermore, merely considering model consensus by performing multiple inter-node communications will lead to a high communication cost. Therefore, the existing works do not jointly consider communication efficiency and model consensus on decentralized learning system.

Motivated by the above observation, we propose a general framework, termed DFL, where each node performs multiple local updates and multiple inter-node communications in a round. DFL can balance communication efficiency and model consensus by collaboratively allocating computing and communication steps. We illustrate the effectiveness of multiple inter-node communications in model consensus as follows.

Without loss of generality, we set a simplified decentralized network for illustration. As shown in Fig. 2, the model contains five nodes which form a ring topology. After local updates, all nodes get their initial parameters $\mathbf{w}_1, ..., \mathbf{w}_5$. Fig. 3 shows the evolution trajectory of coefficients variance of initial parameters with communication step at Node 3. We find that the variance monotonically decreases with communication steps, which means more communications better approach the average model and enhance the model consensus. A theoretical explanation will be provided in Proposition 1.

B. DFL

The DFL learning strategy consists of two stages, which are local update and inter-node communication (communication with neighboring nodes), respectively. Firstly, in local update
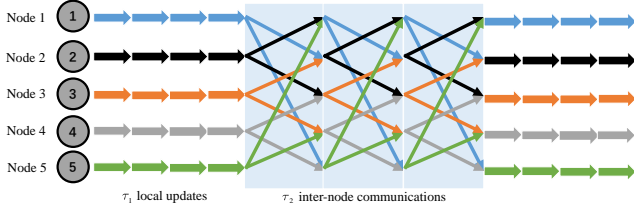
Fig. 4: Illustration of DFL learning strategy. In the decentralized model with 5 nodes in Fig. 2, DFL performs τ_2 inter-node communication steps after τ_1 local update steps.

stage, each node computes the current gradient of local loss function and updates the model parameter by performing SGD multiple times in parallel. After finishing local update, each node performs multiple inter-node communications. In inter-node communication stage, each node sends its local model parameters to the nodes connected with it, and receives the parameters from its neighboring nodes. Then model averaging based on the received parameters is performed by the node to obtain the updated parameters. The whole learning process is the alternation of the two stages.

We define $\mathbf{C} \in \mathbb{R}^{N \times N}$ as the confusion matrix which captures the network topology, and it is doubly stochastic, i.e., $\mathbf{C1} = \mathbf{1}, \mathbf{C}^\top = \mathbf{C}$. The element c_{ji} denotes the contribution of node j in model averaging at node i. We let DFL perform τ_1 local updates in parallel at each node during the local updates stage. After τ_1 local updates, each node needs to perform τ_2 inter-node communications. We call τ_1 *the computation frequency* and τ_2 *the communication frequency* in the sequel. In Fig. 4, we illustrate the learning strategy of DFL. Then we define $\tau \triangleq \tau_1 + \tau_2$ in DFL framework and we name the combination of τ_1 local updates and τ_2 inter-node communications as an iteration *round*. Thus the k-th iteration round is $[(k-1)\tau, k\tau)$ for $k = 1, 2, \ldots$. We divide an iteration round into $[(k-1)\tau, (k-1)\tau + \tau_1)$ and $[(k-1)\tau + \tau_1, k\tau)$ and define $[k]_1 \triangleq [(k-1)\tau, (k-1)\tau + \tau_1)$ and $[k]_2 \triangleq [(k-1)\tau + \tau_1, k\tau)$ for convenience. Thus round $[k]_1$ and $[k]_2$ correspond to the iterative period of local update and inter-node communication, respectively. We use $\mathbf{w}_t^{(i)}, g(\mathbf{w}_t^{(i)})$ to denote the model parameter and the gradient of node i at the t-th iteration, respectively.

The learning strategy of DFL can be described as follows.

Local update: When $t \in [k]_1$, local update is performed at each node in parallel. At node i, the updating rule can be expressed as

$$\mathbf{w}_{t+1}^{(i)} = \mathbf{w}_t^{(i)} - \eta g(\mathbf{w}_t^{(i)}),$$

where η is the learning rate. Note that each node performs SGD individually based on its local dataset.

Inter-node communication: When $t \in [k]_2$, inter-node communication is performed. Each node communicates with the connected nodes to exchange model parameters. After the parameters from all the connected nodes are received, model averaging is performed by

$$\mathbf{w}_{t+1}^{(i)} = \sum_{j=1}^{N} c_{ji} \mathbf{w}_t^{(j)}.$$

Algorithm 1 DFL

Input:

The learning rate η, the total number of iterations T, computation frequency τ_1 and communication frequency τ_2 in an iteration round

1: Set the initial value of $\mathbf{X}_0$.
2: **for** $t = 1, 2, \ldots, K\tau$ **do**
3:    **if** $t \in [k]_1$ where $k = 1, \ldots, K$ **then**
4:      $\mathbf{X}_{t+1} = \mathbf{X}_t - \eta \mathbf{G}_t$ *local update*
5:    **else**
6:      $\mathbf{X}_{t+1} = \mathbf{X}_t \mathbf{C}$ *inter-node communication*
7:    **end if**
8: **end for**
9: **for** $t = K\tau + 1, \ldots, T$ **do**
10:    $\mathbf{X}_{t+1} = \mathbf{X}_t - \eta \mathbf{G}_t$
11: **end for**

For convenience, we reunite the learning strategy into matrix form. We first define *the model parameter matrix* $\mathbf{X}_t$ and *the gradient matrix* $\mathbf{G}_t \in \mathbb{R}^{d \times N}$ as follows:

$$\mathbf{X}_t = \left[\mathbf{w}_t^{(1)}, \mathbf{w}_t^{(2)} \ldots, \mathbf{w}_t^{(N)} \right],$$
$$\mathbf{G}_t = \left[g(\mathbf{w}_t^{(1)}), \ldots, g(\mathbf{w}_t^{(N)}) \right].$$

Therefore, the learning strategy in matrix form can be rewritten as follows.

Local update: $\mathbf{X}_{t+1} = \mathbf{X}_t - \eta \mathbf{G}_t$ for $t \in [k]_1$.

Inter-node communication: $\mathbf{X}_{t+1} = \mathbf{X}_t \mathbf{C}$ for $t \in [k]_2$.

We further combine the two update rules into one to facilitate the following theoretical analysis. Before describing the update rule of DFL, we define two time-varying matrices $\mathbf{C}_t, \mathbf{G}_t'$ as follows:

$$\mathbf{C}_t \triangleq \begin{cases} \mathbf{I}, & t \in [k]_1 \\ \mathbf{C}, & t \in [k]_2, \end{cases} \tag{3}$$

$$\mathbf{G}_t' \triangleq \begin{cases} \mathbf{G}_t, & t \in [k]_1 \\ \mathbf{0}, & t \in [k]_2, \end{cases} \tag{4}$$

where $\mathbf{I} \in \mathbb{R}^{N \times N}$ is the identity matrix and $\mathbf{0} \in \mathbb{R}^{d \times N}$ is the zero matrix. According to the definitions of (3) and (4), **the global learning strategy** of DFL can be described as

$$\mathbf{X}_{t+1} = (\mathbf{X}_t - \eta \mathbf{G}_t') \mathbf{C}_t. \tag{5}$$

Assuming that all nodes have limited communication and computation resources, we use a finite number T to denote the total iteration rounds. Then the corresponding number of the iteration round is denoted by K with $K = \lfloor \frac{T}{\tau} \rfloor^1$. We present the pseudo-code of DFL in Algorithm 1.

A comparison among different distributed SGD methods is shown in Table I. From this table, we can find that DFL is beneficial for alleviating local drift caused by inter-node communication.

---

[1] Note that we set the condition $T - K\tau \leq \tau_1$ in this paper for simplified analysis. The other case $T - K\tau > \tau_1$ can be similarly treated based on the former.

TABLE I: Comparison of different distributed SGD methods

Method	Steps in a round (local update steps, inter-node communication steps)	Local drift due to local computation	Local drift due to inter-node communication	Central server
FL [3]	$(\tau, \text{Null})^1$	High	Null[2]	Required
D-SGD [30]	$(1, 1)$	Low	High	Not needed
C-SGD [31]	$(\tau, 1)$	High	High	Not needed
DFL	(τ_1, τ_2)	High	Low	Not needed

[1] In FL, there is no inter-node communication. A central parameter server is used for collecting local models.
[2] FL achieves model consensus by taking a weighted average of all local models. Thus, local drift from inter-node communication dose not exist.
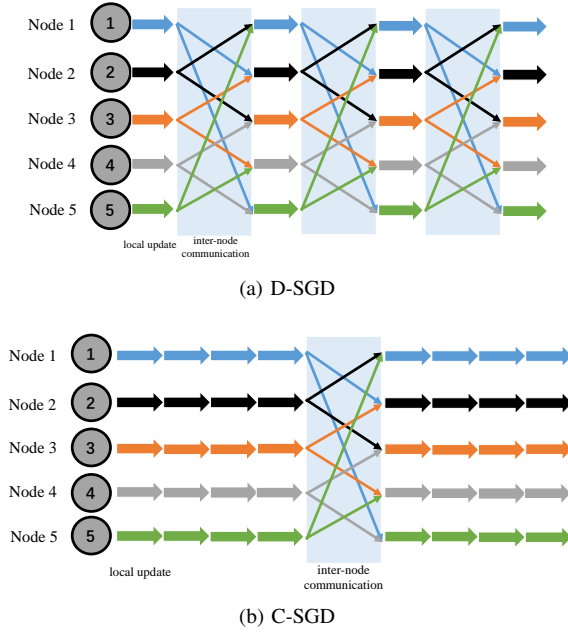

(a) D-SGD

(b) C-SGD

Fig. 5: Learning strategies of D-SGD and C-SGD

C. Special Cases of DFL

When performing inter-node communication only once, DFL can degenerate into two special cases, which are D-SGD and C-SGD, respectively.

1) D-SGD: D-SGD algorithm was studied and developed for solving the decentralized learning problem without aggregating local models at a central server [18] [30]. The network topology of D-SGD is captured by a confusion matrix $\mathbf{C} \in \mathbb{R}^{N \times N}$ whose (j, i)-th element c_{ji} represents the contribution of node j in model averaging at node i. As shown in Fig. 5a, nodes of D-SGD execute *local update* and *inter-node communication* once alternately. The learning strategy of D-SGD consists of the two steps as follows.

Inter-node communication: For each node i, the neighboring nodes that are connected with node i send their current model parameters to node i at the t-th iteration. Then node i takes a weighted average of the received parameters and the local parameter by

$$\mathbf{w}_{t+\frac{1}{2}}^{(i)} = \sum_{j=1}^{N} c_{ji}\mathbf{w}_t^{(j)}, \qquad (6)$$

where $\mathbf{w}_{t+\frac{1}{2}}^{(i)}$ denotes the parameter of model averaging after inter-node communication. Note that only the nodes connected with node i contribute to the model average of node i. If node j $(j = 1, ..., N)$ is not connected with node i, then $c_{ji} = 0$; otherwise, $c_{ji} > 0$. Of course, $c_{ii} > 0$ is necessary.

Local update: After inter-node communication, each node i performs a local update as follows:

$$\mathbf{w}_{t+1}^{(i)} = \mathbf{w}_{t+\frac{1}{2}}^{(i)} - \eta g(\mathbf{w}_t^{(i)}), \qquad (7)$$

where η is the learning rate. The local update follows the SGD method.

At iteration t, *the model parameter matrix* and *the gradient matrix* are denoted by $\mathbf{X}_t, \mathbf{G}_t \in \mathbb{R}^{d \times N}$, respectively. Then according to (6) and (7), we have **the global learning strategy** of D-SGD by combining the two steps as follows:

$$\mathbf{X}_{t+1} = \mathbf{X}_t \mathbf{C} - \eta \mathbf{G}_t. \qquad (8)$$

2) C-SGD: The work [31] proposed C-SGD which allows nodes to perform multiple local updates and inter-node communication once to improve communication efficiency. At t-th iteration, all nodes have different model parameters $\mathbf{w}_t^{(1)}, ..., \mathbf{w}_t^{(N)} \in \mathbb{R}^d$. The network topology of C-SGD is captured by the confusion matrix $\mathbf{C} \in \mathbb{R}^{N \times N}$ again. The learning strategy of C-SGD has similar steps as D-SGD. As shown in Fig. 5b, C-SGD performs an inter-node communication after τ local updates. The learning strategy of C-SGD is presented as follows.

Local update: For each node i, when the iteration index t satisfies $t \bmod \tau \neq 0$, a local update is performed according to

$$\mathbf{w}_{t+1}^{(i)} = \mathbf{w}_t^{(i)} - \eta g(\mathbf{w}_t^{(i)}), \qquad (9)$$

where η is the learning rate.

Inter-node communication: After the stage of local update, each node performs inter-node communication with its connected nodes. At the t-th iteration satisfying $t \bmod \tau = 0$, node i receives the model parameters from its neighbors, and computes a weighted averaging as follows:

$$\mathbf{w}_{t+1}^{(i)} = \sum_{j=1}^{N} c_{ji}(\mathbf{w}_t^{(i)} - \eta g(\mathbf{w}_t^{(i)})). \qquad (10)$$

Then we present the matrix format of the C-SGD learning process. The time-varying confusion matrix $\mathbf{C}_t \in \mathbb{R}^{N \times N}$ is defined as

$$\mathbf{C}_t = \begin{cases} \mathbf{C}, & t \bmod \tau = 0 \\ \mathbf{I}, & \text{otherwise,} \end{cases}$$

where $\mathbf{I} \in \mathbb{R}^{N \times N}$ is the identity matrix, which means there is no inter-node communication during the local updates stage. Then, at the t-th iteration , *the model parameter matrix* and *the gradient matrix* are denoted by $\mathbf{X}_t, \mathbf{G}_t \in \mathbb{R}^{d \times N}$, respectively. Thus, **the global learning strategy** of C-SGD can be rewritten as

$$\mathbf{X}_{t+1} = (\mathbf{X}_t - \eta \mathbf{G}_t)\mathbf{C}_t. \tag{11}$$

Note that the order of inter-node communication and local update can be reversed in D-SGD and C-SGD, resulting in difference of the global learning strategy of the two cases as shown in (8) and (11). We call (8) communicate-then-compute and (11) compute-then-communicate. In fact, communicate-then-compute and compute-then-communicate have equivalent convergence behavior, as revealed by our analysis in Section III-C3.

3) Convergence of D-SGD and C-SGD: We present the learning strategies of the two methods as follows:

$$\text{D-SGD:} \quad \mathbf{X}_{t+1} = \mathbf{X}_t \mathbf{C} - \eta \mathbf{G}_t, \tag{12}$$

$$\text{C-SGD:} \quad \mathbf{X}_{t+1} = (\mathbf{X}_t - \eta \mathbf{G}_t)\mathbf{C}, \tag{13}$$

where (12) and (13) denote the learning strategies of communicate-then-compute and compute-then-communicate, respectively.

For facilitating the analysis, we multiply $\mathbf{1}/N$ on both sides of (12) and (13) to obtain the format of model averaging. So we have

$$\mathbf{X}_{t+1}\frac{\mathbf{1}}{N} = \mathbf{X}_t\mathbf{C}\frac{\mathbf{1}}{N} - \eta\mathbf{G}_t\frac{\mathbf{1}}{N}, \tag{14}$$

$$\mathbf{X}_{t+1}\frac{\mathbf{1}}{N} = (\mathbf{X}_t - \eta\mathbf{G}_t)\mathbf{C}\frac{\mathbf{1}}{N}. \tag{15}$$

Because the confusion matrix $\mathbf{C}$ is doubly stochastic, we have $\mathbf{C1} = \mathbf{1}$. Therefore, we can rewrite (14) and (15) as

$$\text{D-SGD:} \quad \mathbf{u}_{t+1} = \mathbf{u}_t - \eta\left[\frac{1}{N}\sum_{i=1}^{N} g(\mathbf{w}_t^{(i)})\right],$$

$$\text{C-SGD:} \quad \mathbf{u}_{t+1} = \mathbf{u}_t - \eta\left[\frac{1}{N}\sum_{i=1}^{N} g(\mathbf{w}_t^{(i)})\right],$$

respectively, where we define $\mathbf{u}_t \triangleq \mathbf{X}_t\frac{\mathbf{1}}{N}$. Consequently, we find out that the learning strategies of communicate-then-compute and compute-then-communicate have the same update rules based on the averaged model $\mathbf{u}_t$. Thus, we can conclude that communicate-then-compute and compute-then-communicate lead to an equivalent convergence bound.

IV. CONVERGENCE ANALYSIS

In this section, we analyze the convergence of the proposed algorithm in DFL framework and further study how the choice of τ_1 and τ_2 impacts the convergence bound.

A. Preliminaries

To facilitate the analysis, we make the following assumptions.

Assumption 1. *we assume the following conditions:*
1) $F_i(\mathbf{w})$ *is L-smooth, i.e.,* $\|\nabla F_i(\mathbf{x}) - \nabla F_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ *for some $L > 0$ and any $\mathbf{x}$, $\mathbf{y}$, i. Hence, we have $F(\mathbf{w})$ is L-smooth from the triangle inequality.;*
2) $F_i(\mathbf{w})$ *is μ-strong;*
3) $F(\mathbf{w})$ *has a lower bound, i.e.,* $F(\mathbf{w}) \geq F_{\inf}$ *for $F_{\inf} > 0$;*
4) *Gradient estimation is unbiased for stochastic mini-batch sampling, i.e.,* $\mathbb{E}_{\xi|\mathbf{w}}[g(\mathbf{w})] = \nabla F(\mathbf{w})$;
5) *Gradient estimation has a bounded variance, i.e.,* $\mathbb{E}_{\xi|\mathbf{w}}\|g(\mathbf{w}) - \nabla F(\mathbf{w})\|^2 \leq \sigma^2$ *where $\sigma^2 > 0$. Furthermore, for any node i,* $\mathbb{E}_{\xi|\mathbf{w}}\|g(\mathbf{w}^{(i)}) - \nabla F_i(\mathbf{w})\|^2 \leq \sigma_i^2$ *and* $\mathbb{E}_{\xi|\mathbf{w}}\|g(\mathbf{w}^{(i)})\|^2 \leq G^2$. *We use $\bar{\sigma}^2$ to denote the average of $\{\sigma_i^2\}$, i.e.,*

$$\bar{\sigma}^2 \triangleq \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2;$$

6) $\mathbf{C}$ *is a doubly stochastic matrix, which satisfies* $\mathbf{C1} = \mathbf{1}, \mathbf{C}^\top = \mathbf{C}$. *The largest eigenvalue of $\mathbf{C}$ is always 1 and the others are strictly less than 1, i.e.,* $\max\{|\lambda_2(\mathbf{C})|, |\lambda_N(\mathbf{C})|\} < \lambda_1(\mathbf{C}) = 1$, *where $\lambda_i(\mathbf{C})$ denotes the i-th largest eigenvalue of $\mathbf{C}$. For convenience, we define*

$$\beta \triangleq \|\mathbf{I} - \mathbf{C}\|_2 \in [0, 2],$$

which is a similar way to analyze communication-compressed D-SGD in [34].

B. Update of Model Averaging

For convenience of analysis, we transform the update rule of DFL in (5) into model averaging. By multiplying $\mathbf{1}/N$ on both sides of (5), we obtain

$$\mathbf{X}_{t+1}\frac{\mathbf{1}}{N} = \mathbf{X}_t\frac{\mathbf{1}}{N} - \eta\mathbf{G}_t'\frac{\mathbf{1}}{N}, \tag{16}$$

where $\mathbf{C}_t$ is eliminated because of $\mathbf{C}_t\mathbf{1} = \mathbf{1}$ from the definition of $\mathbf{C}_t$ in (3). Thus, we can obtain

$$\mathbf{u}_{t+1} = \mathbf{u}_t - \eta\left[\frac{1}{N}\sum_{i=1}^{N} g(\mathbf{w}_t^{(i)})\right] \tag{17}$$

for $t \in [k]_1$ and $\mathbf{u}_{t+1} = \mathbf{u}_t$ for $t \in [k]_2$. This is because SGD algorithm is only performed in local updates. In the remaining analysis, we will concentrate on the convergence of the average model $\mathbf{u}_t$, which is a common way to analyze convergence under stochastic sampling setting of gradient descent [9] [18] [35].

Because the global loss function $F(\mathbf{w})$ can be non-convex in complex learning platform such as Convolutional Neural Network (CNN), SGD may converge to a local minimum or a saddle point. So we use the expectation of the gradient norm average of all iterations as the indicator of convergence. Similar indicators have been used in [36] [37]. The algorithm is convergent if the following condition is satisfied.

Definition 1 (**Algorithm Convergence**). *The algorithm converges to a stationary point if it achieves an ϵ-optimal solution, i.e.,*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{u}_t)\|^2\right] \leq \epsilon. \tag{18}$$

C. Convergence Results

Based on Assumption 1 and Definition 1, we have the following proposition which characterizes the convergence of DFL.

Proposition 1 (**Convergence of DFL**). *For DFL algorithm, let the total iterations be T and the iteration round be $\tau = \tau_1 + \tau_2$, where τ_1 and τ_2 are the computation frequency and the communication frequency, respectively. Based on Assumption 1 and that all local models have the same initial point $\mathbf{u}_1$, if the learning rate η satisfies*

$$\eta L + \frac{\eta^2 L^2 \tau}{1-\zeta^{\tau_2}}\left(\frac{2\tau_1\zeta^{2\tau_2}}{1+\zeta^{\tau_2}}+\frac{2\tau_1\zeta^{\tau_2}}{1-\zeta^{\tau_2}}+\tau-1\right)\leq 1, \tag{19}$$

where $\zeta = \max\{|\lambda_2(\mathbf{C})|, |\lambda_N(\mathbf{C})|\}$ with $0 \leq \zeta \leq 1$, then the expectation of the gradient norm average after T iterations is bounded as follows:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{u}_t)\|^2\right] \leq \underbrace{\frac{2[F(\mathbf{u}_1)-F_{\inf}]}{\eta T}+\frac{\eta L\sigma^2}{N}}_{synchronous\ SGD}+$$

$$\underbrace{2\eta^2 L^2\sigma^2\left(\frac{\tau_1}{1-\zeta^{2\tau_2}}-1\right)}_{local\ drift} \tag{20}$$

$$\stackrel{T\to\infty}{\longrightarrow} \frac{\eta L\sigma^2}{N}+2\eta^2 L^2\sigma^2\left(\frac{\tau_1}{1-\zeta^{2\tau_2}}-1\right), \tag{21}$$

where $\mathbf{u}_t = \mathbf{X}_t\frac{1}{N}$.

Proof. The complete proof is presented in Appendix A. We present the proof outline as follows.

From Definition 1, we first derive an upper bound of $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{u}_t)\|^2\right]$, which is give by Lemma 1 in Appendix A as follows:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{u}_t)\|^2\right] \leq \underbrace{\frac{2[F(\mathbf{w}_1)-F_{\inf}]}{\eta T}+\frac{\eta L\sigma^2}{N}}_{synchronous\ SGD}+$$

$$\underbrace{\frac{L^2}{T}\sum_{t=1}^{T}\frac{\mathbb{E}\|\mathbf{X}_t(\mathbf{I}-\mathbf{J})\|_{\mathrm{F}}^2}{N}}_{local\ drift}. \tag{22}$$

Then considering the second term of (22), we can find an upper bound of $\mathbb{E}\|\mathbf{X}_t(\mathbf{I}-\mathbf{J})\|_{\mathrm{F}}^2$, i.e.,

$$\mathbb{E}\|\mathbf{X}_t(\mathbf{I}-\mathbf{J})\|_{\mathrm{F}}^2$$

$$\leq 2\eta^2\mathbb{E}\underbrace{\left\|\sum_{r=0}^{j}(\mathbf{Y}_r-\mathbf{Q}_r)(\mathbf{C}^{(j-r)\tau_2}-\mathbf{J})\right\|_{\mathrm{F}}^2}_{A_1}$$

$$+ 2\eta^2\mathbb{E}\underbrace{\left\|\sum_{r=0}^{j}\mathbf{Q}_r(\mathbf{C}^{(j-r)\tau_2}-\mathbf{J})\right\|_{\mathrm{F}}^2}_{A_2}.$$

Finally, we decompose $\mathbb{E}\|\mathbf{X}_t(\mathbf{I}-\mathbf{J})\|_{\mathrm{F}}^2$ into two parts, denoted as A_1 and A_2, respectively. For A_1, we have

$$\sum_{t=1}^{T}A_1 \leq$$

$$2\eta^2 N\sigma^2 T\left(\frac{\tau_1}{1-\zeta^{2\tau_2}}-1\right)+\frac{2\eta^2\beta\tau}{1-\zeta^{2\tau_2}}\sum_{r=0}^{\frac{T}{\tau}-1}\sum_{s=r\tau+1}^{r\tau+\tau_1}\|\nabla F(\mathbf{X}_s)\|_{\mathrm{F}}^2;$$

and for A_2, we have

$$\sum_{t=1}^{T}A_2 \leq$$

$$\frac{\eta^2\tau}{1-\zeta^{\tau_2}}\left(\frac{2\tau_1\zeta^{2\tau_2}}{1+\zeta^{\tau_2}}+\frac{2\tau_1\zeta^{\tau_2}}{1-\zeta^{\tau_2}}+\tau-1\right)\sum_{r=0}^{\frac{T}{\tau}-1}\sum_{s=r\tau+1}^{r\tau+\tau_1}\mathbb{E}\|\nabla F(\mathbf{X}_s)\|_{\mathrm{F}}^2.$$

From the upper bounds of A_1 and A_2, we finish the proof of Proposition 1.

Note that the assumption of μ-strong convexity in Assumption 1 is not necessary for the convergence bound in (20). $\square$

Focusing on the right side of the inequality (20), we find out that the convergence upper bound of DFL consists of two parts as indicated by the under braces. The first two terms are the original error bound led by synchronous SGD [38], where one local update follows one model averaging covering all nodes, not just the neighbors. The synchronous SGD also corresponds to a special DFL case with $\tau_1 = 1, \tau_2 \to \infty$. The last term is the error bound brought by local drift, which reflects the effect of performing multiple local updates and insufficient model averaging. The consensus model can not be acquired due to finite communication frequency τ_2. Focusing on (21), we can find out that DFL algorithm converges to a stable and static point as $T \to \infty$, regardless of network topology. The network topology parameter given by ζ and the frequencies of computation and communication (i.e., τ_1 and τ_2, respectively) will affect the final convergence upper bound.

We further analyze how the computation frequency τ_1, the communication frequency τ_2 and network topology parameter ζ affect the convergence of DFL.

Remark 1 (**Impact of τ_1, τ_2**). Focusing on the last term in (20), we can find how the computation frequency τ_1 and the communication frequency τ_2 affect the convergence of DFL. Specifically, the bound increases with τ_1 and decreases with τ_2 monotonically. The former is because more local updates lead

to larger difference between the local models and the previous average model over neighboring nodes. Thus the local drift is intensified with τ_1. The latter is because more inter-node communications can facilitate model consensus and make local model approach the ideal average model. Thus the local drift is mitigated by increasing τ_2. DFL with $\tau_2 = 1$ is equivalent to C-SGD, which has a worse convergence rate compared to DFL with $\tau_2 > 1$. The local drift will tend to zero based on the ideal condition that $\tau_1 = 1$ and $\tau_2 \to \infty$, which is the synchronous SGD. In contrast, the local drift will become the worst when τ_1 is large and $\tau_2 = 1$.

Corollary 1. *If $\tau_1 = 1, \tau_2 \to \infty$, DFL achieves the optimal convergence bound without local drift which is*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{u}_t)\|^2\right] \leq \frac{2[F(\mathbf{u}_1) - F_{\inf}]}{\eta T} + \frac{\eta L \sigma^2}{N}. \quad (23)$$

Proof. It directly follows from (20). $\qquad\square$

Remark 2 (Impact of ζ). From (20), the local drift will monotonically increase with ζ. As we defined, ζ denotes the second largest eigenvalue of $\mathbf{C}$ and it reflects the exchange rate of all local models during a single inter-node communication step. Thus a larger ζ means a sparser matrix, resulting to a worse model averaging and aggravating the local drift. When each node can not communicate with any other nodes in the network, the confusion matrix $\mathbf{C} = \mathbf{I}$ and $\zeta = 1$. There does not exist model averaging and it will cause the worst local drift. When each node can communicate with all nodes in the network, the confusion matrix $\mathbf{C} = \mathbf{J}$ and $\zeta = 0$. Model consensus is achieved and the local drift brought by insufficient model averaging is eliminated.

Corollary 2. *If $\zeta = 0, \mathbf{C} = \mathbf{J}$, DFL achieved a convergence bound without local drift caused by insufficient model averaging, which is expressed as*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{u}_t)\|^2\right] \leq \frac{2[F(\mathbf{u}_1) - F_{\inf}]}{\eta T} + \frac{\eta L \sigma^2}{N} + $$
$$2\eta^2 L^2 \sigma^2 (\tau_1 - 1). \quad (24)$$

Proof. It directly follows from (20). $\qquad\square$

V. COMMUNICATION-EFFICIENT DFL BASED ON COMPRESSED COMMUNICATION

In this section, we further study DFL based on a compressed communication theme to reduce the communication overhead when considering multiple inter-node communication steps in a round. The learning strategy of DFL with compressed communication is introduced and the algorithm is developed. Then we present the convergence analysis of DFL with compressed communication.

A. The Learning Strategy of C-DFL

In order to further improve communication efficiency of DFL, we apply the compressed inter-node communication scheme to DFL algorithm. Focusing on the inter-node communication stage, for each node i, the gossip type algorithm

[39] can exchange the model parameters by iterations of the form

$$\mathbf{w}_{t+1}^{(i)} = \mathbf{w}_t^{(i)} + \gamma \sum_{j=1}^{N} c_{ji}\Delta_t^{(ji)}, \quad (25)$$

where $\gamma \in (0, 1]$ denotes the consensus step size and $\Delta_t^{(ji)} \in \mathbb{R}^d$ denotes a vector sent from node j to node i in the t-th iteration. Note that if $\gamma = 1$ and $\Delta_t^{ji} = \mathbf{w}_t^{(j)} - \mathbf{w}_t^{(i)}$, we have $\mathbf{w}_{t+1}^{(i)} = \sum_{j=1}^{N} c_{ji}\mathbf{w}_t^{j}$, which is consistent with the inter-node communication rule of DFL. We apply the CHOCO-G gossip scheme to DFL with compressed communication to remove the compression noise as $t \to \infty$ and guaranteeing model average-consensus [34]. The scheme is presented as

$$\Delta_t^{(ji)} = \hat{\mathbf{w}}_t^{(j)} - \hat{\mathbf{w}}_t^{(i)}, \quad (26)$$
$$\hat{\mathbf{w}}_{t+1}^{(j)} = \hat{\mathbf{w}}_t^{(j)} + Q(\mathbf{w}_{t+1}^{(j)} - \hat{\mathbf{w}}_t^{(j)}), \quad (27)$$

where Q is a compression operator and $\hat{\mathbf{w}}_t^{(i)} \in \mathbb{R}^d$ denotes additional variables which can be obtained and stored by all neighboring nodes (including node i itself) of node i.

We introduce some examples of compression strategies [34].
- *Sparsification*: We randomly select k out of d coordinates (rand$_k$), or the k coordinates with highest magnitude values (top$_k$), which gives $\delta = \frac{k}{d}$.
- *Rescaled unbiased estimators*: Assuming $\mathbb{E}_Q Q(\mathbf{x}) = \mathbf{x}$, $\forall \mathbf{x} \in \mathbb{R}^d$ and $\mathbb{E}_Q \|Q(\mathbf{x})\|^2 \leq c\|\mathbf{x}\|^2$, we then have $Q'(\mathbf{x}) = \frac{1}{c}Q(\mathbf{x})$ with $\delta = \frac{1}{c}$.
- *Randomized gossip*: Supposing $Q(\mathbf{x}) = \mathbf{x}$ with probability $p \in (0, 1]$ and $Q(\mathbf{x}) = 0$ otherwise, we have $\delta = p$.
- *Random quantization*: For the quantization levels $s \in \mathbb{N}_+$ and $c = (1 + \min\{d/s^2, \sqrt{d}/s\})$, the quantization operator can be formulated as

$$\text{qsgd}_s(x) = \frac{\text{sign}(\mathbf{x}) \cdot \|\mathbf{x}\|}{sc} \cdot \left\lfloor s\frac{|x|}{\|\mathbf{x}\|} + \xi \right\rfloor,$$

where the random variable $\xi \sim_{\text{u.a.r.}} [0, 1]^d$ satisfies (28) with $\delta = \frac{1}{c}$.

Note that δ is the compression rate, which reflects the compression quality for the compression operator Q.

We apply the CHOCO-G scheme to inter-node communication of DFL for $t \in [k]_2$, and local update for $t \in [k]_1$ remains unchanged. The proposed communication-efficient algorithm is named DFL with compressed communication (C-DFL). C-DFL is presented in Algorithm 2. Note that in C-DFL, we set $T = K\tau$ for convenience. According to Algorithm 2, local update step with SGD is in line 4, iterative update of inter-node communication step is in line 6, compression is performed in line 7, the exchange of compressed information is in line 10, and the iteration of additional variables is in line 11.

Because multiple inter-node communications with compression and multiple local updates are performed in a round, C-DFL is a more general analytical framework compared with CHOCO-SGD proposed in [34].

B. Convergence Analysis of C-DFL

In this subsection, we analyze the convergence of Algorithm 2. We make an assumption about the compression operator Q to measure the compression quality of different methods.

Algorithm 2 C-DFL

Input:

The total number of iterations T, computation frequency τ_1 and communication frequency τ_2 in an iteration round, initial values $\mathbf{w}_0^{(i)}$ for all nodes, the learning rate η_k changed with k, the consensus step size γ

1: Set the initial value $\hat{\mathbf{w}}_0^i = \mathbf{0} \ \forall i$.
2: **for** $t = 1, 2, ..., K\tau$ in parallel for all nodes i **do**
3:    **if** $t \in [k]_1$ where $k = 1, ..., K$ **then**
4:       $\mathbf{w}_{t+1}^{(i)} = \mathbf{w}_t^{(i)} - \eta_k g(\mathbf{w}_t^{(i)})$
5:    **else**
6:       $\mathbf{w}_{t+1}^{(i)} = \mathbf{w}_t^{(i)} + \gamma \sum_{j=1}^N c_{ji}(\hat{\mathbf{w}}_t^{(j)} - \hat{\mathbf{w}}_t^{(i)})$
7:       $\mathbf{q}_t^{(i)} = Q(\mathbf{w}_{t+1}^{(i)} - \hat{\mathbf{w}}_t^{(i)})$
8:    **end if**
9:    **for** neighboring nodes j of node i (including node i itself) **do**
10:       Send $\mathbf{q}_t^{(i)}$ and receive $\mathbf{q}_t^{(j)}$
11:       $\hat{\mathbf{w}}_{t+1}^{(j)} = \hat{\mathbf{w}}_t^{(j)} + \mathbf{q}_t^{(j)}$
12:    **end for**
13: **end for**

Assumption 2 (Compression operator)**.** *We assume that $\forall \mathbf{x} \in \mathbb{R}^d$, the compression operator Q: $\mathbb{R}^d \to \mathbb{R}^d$ satisfies*

$$\mathbb{E}_Q \|Q(\mathbf{x}) - \mathbf{x}\|^2 \le (1 - \delta)\|\mathbf{x}\|^2, \tag{28}$$

for the compression ratio $\delta > 0$, where $\mathbb{E}_Q$ denotes the expectation about the internal randomness of operator Q.

We define $\mathbf{u}^{(k)} \triangleq \mathbf{u}_{k\tau}$, which is the average of model parameters after k iteration rounds. With Assumption 1 qualifying the loss function $F_i(\cdot)$ and Assumption 2 bounding the discrepancy of compression, we obtain the following proposition which states the convergence of C-DFL.

Proposition 2 (**Convergence of C-DFL**)**.** *Under Assumption 1 and 2, if the learning rate $\eta_k = \frac{4}{\mu(a+k)}$ for parameter $a \ge 16\kappa$, consensus step size $\gamma = \frac{\rho^2 \delta}{16\rho + \rho^2 + 4\beta^2 + 2\rho\beta^2 - 8\rho\delta}$, and the spectral gap $\rho = 1 - \max\{|\lambda_2(\mathbf{C})|, |\lambda_N(\mathbf{C})|\} \in (0, 1]$, then $\mathbb{E}[F(\mathbf{w}_{avg}^{(T)})]$ is bounded as follows:*

$$\mathbb{E}[F(\mathbf{w}_{avg}^{(T)})] - F^* = \mathcal{O}\left(\frac{\tau_1 \bar{\sigma}^2}{\mu K N}\right) + \mathcal{O}\left(\frac{\kappa \tau_1^4 G^2 (1-p)^{\tau_2}}{\mu K^2 [1 - (1+\theta)(1-p)^{\tau_2}]}\right)$$
$$+ \mathcal{O}\left(\frac{\kappa^3 G^2 \tau_1}{K^3 \mu}\right), \tag{29}$$

where $\mathbf{w}_{avg}^{(T)} = \frac{1}{S_K} \sum_{k=0}^{K-1} w_k \mathbf{u}^{(k)}$ for $w_k = (a+k)^2$, $S_K = \sum_{k=0}^{K-1} w_k \ge \frac{1}{3} K^3$, $\theta < \frac{p}{1-p}$ with $p = \frac{\rho^2 \delta}{82}$, and $\kappa = \frac{L}{\mu}$. Note that F^* is the global minimum of $F(\cdot)$.

Proof. The complete proof is presented in Appendix B. We give the proof outline as follows.

We first have

$$\|\mathbf{u}^{(k+1)} - \mathbf{u}^*\|^2$$
$$= \left\| \mathbf{u}^{(k)} - \mathbf{u}^* - \eta_k \sum_{j=0}^{\tau_1 - 1} \frac{1}{N} \sum_{q=1}^N g(\mathbf{w}_{k\tau+j}^{(q)}) \right\|^2$$
$$= \left\| \mathbf{u}^{(k)} - \mathbf{u}^* - \sum_{j=0}^{\tau_1 - 1} \frac{\eta_k}{N} \sum_{i=1}^N \nabla F_i(\mathbf{w}_{k\tau+j}^{(i)}) + \right.$$
$$\left. \sum_{j=0}^{\tau_1 - 1} \frac{\eta_k}{N} \sum_{i=1}^N \nabla F_i(\mathbf{w}_{k\tau+j}^{(i)}) - \sum_{j=0}^{\tau_1 - 1} \frac{\eta_k}{N} \sum_{q=1}^N g(\mathbf{w}_{k\tau+j}^{(q)}) \right\|^2$$
$$\le \left\| \mathbf{u}^{(k)} - \mathbf{u}^* - \sum_{j=0}^{\tau_1 - 1} \frac{\eta_k}{N} \sum_{i=1}^N \nabla F_i(\mathbf{w}_{k\tau+j}^{(i)}) \right\|^2 + \eta_k^2 \tau_1^2 \frac{\bar{\sigma}^2}{N}. \tag{30}$$

Then considering the first term of (30), we obtain

$$\left\| \mathbf{u}^{(k)} - \mathbf{u}^* - \sum_{j=0}^{\tau_1 - 1} \frac{\eta_k}{N} \sum_{i=1}^N \nabla F_i(\mathbf{w}_{k\tau+j}^{(i)}) \right\|^2$$
$$= \|\mathbf{u}^{(k)} - \mathbf{u}^*\|^2 + \eta_k^2 \underbrace{\left\| \sum_{j=0}^{\tau_1 - 1} \frac{1}{N} \sum_{i=1}^N \nabla F_i(\mathbf{w}_{k\tau+j}^{(i)}) \right\|^2}_{T_1}$$
$$- 2\eta_k \underbrace{\left\langle \mathbf{u}^{(k)} - \mathbf{u}^*, \sum_{j=0}^{\tau_1 - 1} \frac{1}{N} \sum_{i=1}^N \nabla F_i(\mathbf{w}_{k\tau+j}^{(i)}) \right\rangle}_{T_2}. \tag{31}$$

Considering T_1 in (31), we have

$$T_1$$
$$\le \frac{2L^2}{N} \tau_1 \sum_{j=0}^{\tau_1 - 1} \sum_{i=1}^N \|\mathbf{w}_{k\tau+j}^{(i)} - \mathbf{u}_{k\tau+j}\|^2 + 4L\tau_1^2 \alpha_k(F(\mathbf{u}^{(k)}) - F^*);$$

and considering T_2 in (31), we obtain

$$-\frac{1}{\eta_k} T_2$$
$$\le -2\tau_1(F(\mathbf{u}^{(k)}) - F^*) + \frac{L+\mu}{N} \sum_{j=0}^{\tau_1 - 1} \sum_{i=1}^N \|\mathbf{u}^{(k)} - \mathbf{w}_{k\tau+j}^{(i)}\|^2$$
$$- \frac{\mu\tau_1}{2} \|\mathbf{u}^{(k)} - \mathbf{u}^*\|^2.$$

We can further get

$$\|\mathbf{u}^{(k)} - \mathbf{w}_{k\tau+j}^{(i)}\|^2$$
$$\le 2\|\mathbf{u}_{k\tau+j} - \mathbf{w}_{k\tau+j}^{(i)}\|^2 + \frac{4\eta_k^2 L^2 j}{N} \sum_{p=0}^{j-1} \sum_{i=1}^N \|\mathbf{w}_{k\tau+p}^{(i)} - \mathbf{u}_{k\tau+p}\|^2 +$$
$$8\eta_k^2 L j^2 \alpha_k(F(\mathbf{u}^{(k)}) - F^*) + \frac{\eta_k^2 j^2 \bar{\sigma}^2}{N}.$$

Then putting everything together, we obtain Lemma 8 in Appendix B, where a upper bound of $\mathbb{E}\|\mathbf{u}_{(k+1)\tau} - \mathbf{u}^*\|^2$ is derived as

$$\mathbb{E}_{\xi_{k\tau}^{(1)},...,\xi_{k\tau}^{(N)}}\|\mathbf{u}^{(k+1)} - \mathbf{u}^*\|^2 \leq \left(1 - \frac{\mu\eta_k\tau_1}{2}\right)\|\mathbf{u}^{(k)} - \mathbf{u}^*\|^2 +$$
$$\eta_k^2\tau_1^2\frac{\bar{\sigma}^2}{N} + \frac{(L+\mu)\eta_k^3\bar{\sigma}^2(\tau_1 - 1)(2\tau_1 - 1)\tau_1}{6N} +$$
$$(F(\mathbf{u}^{(k)}) - F^*)\left[4L\eta_k^2\tau_1^2\alpha_k - 2\eta_k\tau_1 + \frac{4}{3}(L+\mu)\eta_k^3 L\alpha_k(\tau_1 - 1)(2\tau_1 - 1)\tau_1\right]$$
$$+\frac{2L^2\eta_k^2\tau_1}{N}\Phi_k + \frac{2(L+\mu)\eta_k}{N}\Phi_k + \frac{(L+\mu)\eta_k}{N}\sum_{j=0}^{\tau_1-1}\sum_{i=1}^{N}\frac{4\eta_k^2 L^2 j^2}{N}\Phi_{k,j},$$

$$(32)$$

where $\Phi_k = \sum_{j=0}^{\tau_1-1}\sum_{i=1}^{N}\|\mathbf{u}_{k\tau+j} - \mathbf{w}_{k\tau+j}^{(i)}\|^2$ and $\Phi_{k,j} = \sum_{p=0}^{j-1}\sum_{i=1}^{N}\|\mathbf{u}_{k\tau+p} - \mathbf{w}_{k\tau+p}^{(i)}\|^2$. Then upper bounds of Φ_k and $\Phi_{k,j}$ are given in Lemma 11 of Appendix B.

Lemma 12 [35, Lemma 3.4] gives an upper bound about $F(\mathbf{u}_{k\tau})$ because (32) matches the inequality of Lemma 12. Therefore, we can prove the convergence of C-DFL as presented in Lemma 13. Finally, Proposition 2 follows from Lemma 13 using the inequality $\mathbb{E}\mu\|\mathbf{u}_0 - \mathbf{u}^*\| \leq 2G$ derived in [40, Lemma 2]. $\square$

When K and $\bar{\sigma}$ are sufficiently large, the last two terms in (29) are negligible compared to $\mathcal{O}\left(\frac{\tau_1\bar{\sigma}^2}{\mu KN}\right)$. Furthermore, when $\tau_1 = 1$, this dominant term becomes $\mathcal{O}\left(\frac{\bar{\sigma}^2}{\mu KN}\right)$, which is consistent with the result of CHOCO-SGD proposed in [34]. This also shows the linear convergence rate of decentralized SGD methods. Regarding the computation frequency τ_1, it is easy to find that a larger value of τ_1 leads to a greater convergence upper bound. Regarding the communication frequency τ_2, the second term $\mathcal{O}\left(\frac{\kappa\tau_1^4 G^2(1-p)^{\tau_2}}{\mu K^2[1-(1+\theta)(1-p)^{\tau_2}]}\right)$ decreases with τ_2. The result verifies the impact analysis of τ_1 and τ_2 in Remark 1. The second term in (29) decreases with δ, which illustrates that a smaller compression rate lets the convergence get worse when improving the communication efficiency by compression.

VI. SIMULATION AND DISCUSSION

In this section, we build the DFL and C-DFL frameworks with embedding CNN into each node and evaluate the convergence performance of DFL and communication efficiency of C-DFL based on MNIST and CIFAR-10 datasets.

A. Simulation Setup

Based on Torch module of Python environment, we build a decentralized FL framework where each node can only communicate with its connected nodes. Because sparse topologies are able to improve the convergence of DFL significantly, we use ring topology and quasi-ring topology for experiments, and set 10 nodes in the two ring topologies as shown in Fig. 6. For the two topologies, each node updates its model parameter by taking an average of parameters from the connected neighbors and itself. Note that for ring topology, the second largest eigenvalue $\zeta = 0.87$, and for quasi-ring topology, we have
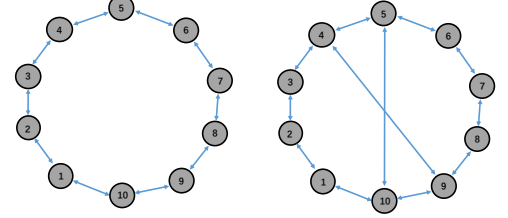

Fig. 6: Ring topology (left) and Quasi-ring topology (right)

$\zeta = 0.85$. We use CNN to train the local model and choose the cross-entropy as the loss function. The specific CNN structure is presented in Appendix C.

In our experiments, we use MNIST [41] and CIFAR-10 [42] dataset for CNN training and testing. Considering actual application scenarios of decentralized FL where different devices have different distribution of data, statistical heterogeneity is a main feature for simulation. So we set that the distribution of the training data samples is non-i.i.d. and the distribution of the testing data is i.i.d.

For experimental setup, SGD is used for local update steps. At the beginning of learning, the same initializations of each model are performed where $\mathbf{w}_0^{(1)} = \mathbf{w}_0^{(2)} = \cdots = \mathbf{w}_0^{(N)}$. In the experiments of using MNIST, the learning rate is set as $\eta = 0.002$. In the experiments of using CIFAR-10, the learning rate is set as $\eta = 0.008$.

B. Simulation Evaluation

In this subsection, we simulate DFL algorithm with CNN and verify the improved convergence of DFL compared with C-SGD. Then we explore the effect of the parameters τ_2, τ_1 and ζ on the convergence performance. Furthermore, we verify the convergence of C-DFL and illustrate the communication efficiency under different compressed communication themes.

1) The accelerated convergence of DFL: We present the accelerated convergence of DFL and use C-SGD as the benchmark, where C-SGD is a special case of DFL with communication frequency $\tau_2 = 1$. We set the fixed computation frequency $\tau_1 = 4$ for DFL and C-SGD algorithms. For ring topology, both DFL and C-SGD are performed based on two CNN models which use MNIST and CIFAR-10 for training and testing, respectively. For quasi-ring topology, DFL and C-SGD algorithms are only based on the CNN model with MNIST. Under the same dataset, the global loss function of the two algorithms are identical.

The curves of training loss and testing accuracy with epoch are presented in Fig. 7. We can see that the training loss and testing accuracy curves converge gradually with iterations based on the two datasets, which validates the effective convergence of DFL algorithms. We can further find out that the train loss and testing accuracy of DFL outperform these of C-SGD under the same iteration. This verifies that DFL improves the convergence compared with C-SGD.

2) Effect of τ_2: As shown in Fig. 7, we can find that under the same iteration, the case of $\tau_2 = 15$ shows the best
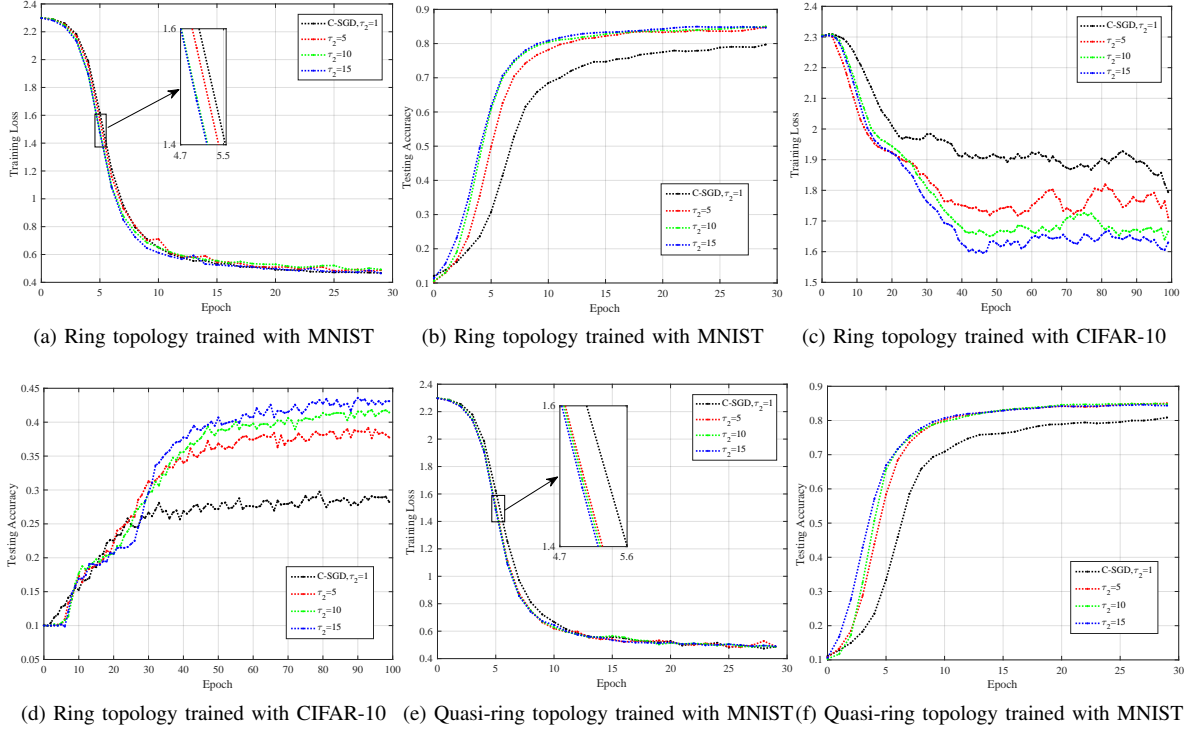
Fig. 7: Experiments on MNIST and CIFAR-10 datasets with CNN under different τ_2. There are 10 nodes for ring topology and quasi-ring topology. We set the computation frequency $\tau_1 = 4$. Ring topology uses both MNIST and CIFAR-10 as the training dataset, and quasi-ring topology is trained with MNIST. For the training on MNIST, we set the learning rate $\eta = 0.002$. For the training on CIFAR-10, we set the learning rate $\eta = 0.008$.
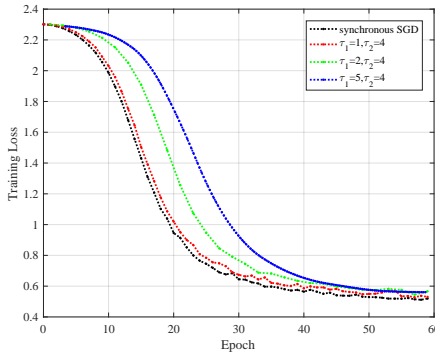

Fig. 8: Training loss curves with different τ_1. The experiment is based on MNIST dataset. There are 10 edge nodes and this experiment is based on ring topology. The learning rate $\eta = 0.002$ and the communication frequency $\tau_2 = 4$.

convergence of training loss and prediction performance and the case of $\tau_2 = 1$ shows the worst one. A larger τ_2 leads to a better convergence whether in training loss or testing accuracy, which verifies the analysis that the convergence upper bound decreases with τ_2 in Remark 1. Thus, more inter-node communications in a round can improve the learning performance of DFL.

3) Effect of τ_1: Based on MNIST dataset, we simulate the training loss curves with different computation frequency τ_1. This experiment is based on ring topology. We use syn-

chronous SGD as the benchmark where the computational frequency is set as $\tau_1 = 1$ and the confusion matrix is set as $\mathbf{C} = \mathbf{J}$ to achieve complete model averaging. The experimental result is presented in Fig. 8. We can find out that under the same iteration, synchronous SGD outperforms DFL, which verifies the Corollary 1 that one local update with sufficient model averaging in a round leads to the optimal convergence of DFL. We can see that as τ_1 goes from 1 to 10, the convergence of training loss becomes much worse. This is because more local updates lead to a worse local drift. Therefore, the simulation results confirm that the convergence bound increases with τ_1 as expressed in Remark 1.

4) Effect of ζ: We also explore the impact of ζ on DFL convergence based on MNIST dataset by simulation. In this experiment, we set the computation frequency $\tau_1 = 2$ and the communication frequency $\tau_2 = 4$. We use DFL with $\zeta = 0$ as the benchmark. From Fig. 9, we can see that DFL with $\zeta = 0$ presents the best convergence performance compared with DFL with $\zeta > 0$. From the perspective of network topology, condition $\zeta = 0$ corresponds to the confusion matrix $\mathbf{C} = \mathbf{J}$ where the network has the densest inter-node connection. So model averaging is achieved, which means there is no local drift caused by insufficient inter-node communication and the convergence of synchronous SGD outperforms that of DFL with $\zeta > 0$. We can further find out that a larger ζ results in a worse convergence of DFL. This is because a larger ζ corresponds to a sparser confusion matrix or network topology. Then more incomplete model averaging is made
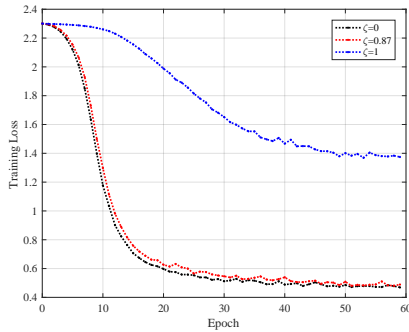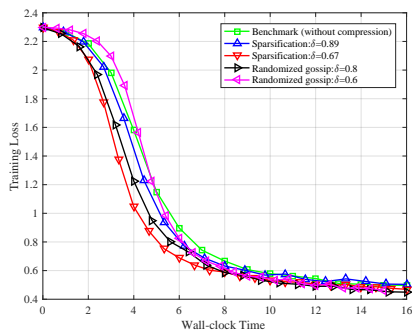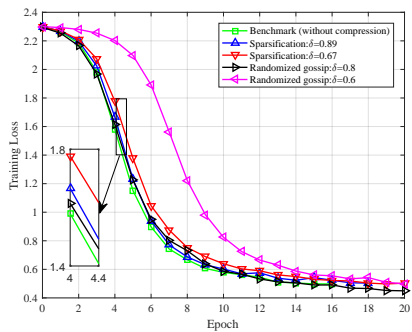
Fig. 9: Training loss curves with different ζ. The experiment is based on MNIST dataset. There are 10 edge nodes. We set the learning rate $\eta = 0.002$, the computation frequency $\tau_1 = 2$ and the communication frequency $\tau_2 = 4$.

(a) Training loss with wall-clock time

(b) Training loss with epoch

Fig. 10: C-DFL training on MNIST with CNN. The ring topology is used with 10 nodes. The computation frequency τ_1 and communication frequency τ_2 are set as 4 and 4, respectively. The consensus step size $\gamma = 1$ and the learning rate η is set as 0.002.

by the sparser topology. Therefore, the simulation of Fig. 9 verifies the impact of ζ in Remark 2.

5) C-DFL: In this simulation, we use the sparsification and randomized gossip as compression operators which are introduced in Section V-A. To show the improved communication efficiency of C-DFL, we use DFL (without compression) as the benchmark. The training is based on MNIST with CNN and the decentralized network is a ring topology. The computation frequency τ_1 and communication frequency τ_2 are set as 4 and 4, respectively. We set the consensus step size $\gamma = 1$ and the

learning rate $\eta = 0.002$.

From Fig. 10, we can see that C-DFL converges under the two compression operators, which are Sparsification and Randomized gossip, respectively. Focusing on Fig. 10(a), we find C-DFL with Sparsification and Randomized gossip converges at a faster rate than DFL. For example, when wall-clock time is 4, we obtain the convergences of Sparsification with $\delta = 0.89, \delta = 0.67$ are increased by 16.9% and 74.6% than DFL, respectively. And the convergence performance of Randomized gossip with $\delta = 0.8$ rises by 50.7% compared with DFL. This shows the communication efficiency of C-DFL because C-DFL transmits fewer bits and accelerates inter-node communication practically. Note that when wall-clock time is smaller than 5, the convergence of Randomized gossip with $\delta = 0.6$ is worse than DFL. We can find Randomized gossip with $\delta = 0.6$ shows the worst convergence in Fig. 10(b). The reason is that because the compression ratio δ is relatively small and too much information is lost in compression, the benefits of reduced communication overhead brought by compression can not offset the inferior convergence performance due to information loss. Therefore, a smaller compression ratio could result in inferior convergence performance about the wall-clock time. Focusing on Fig. 10(b), we can see that convergence of C-DFL is inferior to that of DFL and a smaller compression ratio lets the convergence get worse, which is consistent with the convergence analysis of C-DFL in Proposition 2.

VII. CONCLUSION

In this paper, we have proposed a general decentralized machine learning framework named DFL to jointly consider model consensus and communication efficiency. DFL performs multiple local updates and multiple inter-node communications. Firstly, we have established the strong convergence guarantees for DFL algorithms without the assumption of convex objective. The further analysis about the convergence upper bound of DFL has been made. It shows that the proposed DFL possesses more superior convergence properties compared with C-SGD. Then, we have proposed C-DFL with compressed communication to further improve communication efficiency of DFL. The convergence analysis of C-DFL has been presented and it establishes linear convergence. Finally, based on CNN with training MNIST and CIFAR-10 datasets, experimental simulations have been made. The results have validated the above theoretical analysis of DFL. The effect of C-DFL on improving communication efficiency has been illustrated.

REFERENCES

[1] M. Chiang and T. Zhang. Fog and iot: An overview of research opportunities. *IEEE Internet Things J*, 3(6):854–864, 2016.

[2] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proc. of AISTATS*, pages 1273–1282. PMLR, 2017.

[3] 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.*, 37(6):1205–1221, 2019.

[4] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

[5] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Process. Mag.*, 37(3):50–60, 2020.

[6] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. Federated learning: Strategies for improving communication efficiency. *Proc. NIPS Workshop Private Multi-Party Mach. Learn.*, 2016.

[7] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M. Z. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Y. Ng. Large scale distributed deep networks. In *NIPS*, 2012.

[8] X. Lian, Y. Huang, Y. Li, and J. Liu. Asynchronous parallel stochastic gradient for nonconvex optimization. In *NIPS*, 2015.

[9] H. Yu, S. Yang, and S. Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proc. AAAI Conf. Artif. Intell.*, volume 33, pages 5693–5700, 2019.

[10] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. On the convergence of fedavg on non-iid data. In *Proc. Int. Conf. Learning Representations*, 2020.

[11] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek. Robust and communication-efficient federated learning from non-iid data. *IEEE Trans. Neural Netw. Learn. Syst*, 31(9):3400–3413, 2019.

[12] H. Wang, Z. Kaplan, D. Niu, and B. Li. Optimizing federated learning on non-iid data with reinforcement learning. In *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, pages 1698–1707. IEEE, 2020.

[13] W. Liu, L. Chen, Y. Chen, and W. Zhang. Accelerating federated learning via momentum gradient descent. *IEEE Trans. Parallel Distrib. Syst.*, 31(8):1754–1766, 2020.

[14] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated optimization in heterogeneous networks. *Proc. Conf. Machine Learning and Systems*, 2020.

[15] V. Smith, C.-K. Chiang, M. Sanjabi, and A. Talwalkar. Federated multi-task learning. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 4427–4437, 2017.

[16] F. Ang, L. Chen, N. Zhao, Y. Chen, W. Wang, and F. R. Yu. Robust federated learning with noisy communication. *IEEE Trans. Commun.*, 68(6):3452–3464, 2020.

[17] P. Vanhaesebrouck, A. Bellet, and M. Tommasi. Decentralized collaborative learning of personalized models over networks. In *AISTATS*, pages 509–517. PMLR, 2017.

[18] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 5336–5346, 2017.

[19] C. Hu, J. Jiang, and Z. Wang. Decentralized federated learning: a segmented gossip approach. *Proc. 1st Int. Workshop Fed. Mach. Learn. User Privacy Data Confidentiality (FML)*, 2019.

[20] A. Nedic and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Trans. Autom. Control*, 54(1):48–61, 2009.

[21] T. Wu, K. Yuan, Q. Ling, W. Yin, and A. H. Sayed. Decentralized consensus optimization with asynchrony and delays. *IEEE Trans. Signal Inf. Process. over Netw.*, 4(2):293–307, 2017.

[22] E. Wei and A. Ozdaglar. Distributed alternating direction method of multipliers. In *Proc. IEEE CDC*, pages 5445–5450. IEEE, 2012.

[23] W. Li, Y. Liu, Z. Tian, and Q. Ling. Communication-censored linearized admm for decentralized consensus optimization. *IEEE Trans. Signal Inf. Process. over Netw.*, 6:18–34, 2019.

[24] J. C. Duchi, A. Agarwal, and M. J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Trans. Autom. Control*, 57(3):592–606, 2011.

[25] Z. Yang and W. U. Bajwa. Byrdie: Byzantine-resilient distributed coordinate descent for decentralized learning. *IEEE Trans. Signal Inf. Process. over Netw.*, 5(4):611–627, 2019.

[26] J. Zhang, Q. Ling, and A. M. So. A newton tracking algorithm with exact linear convergence for decentralized consensus optimization. *IEEE Trans. Signal Inf. Process. over Netw.*, 2021.

[27] Q. Lü, X. Liao, H. Li, and T. Huang. A computation-efficient decentralized algorithm for composite constrained optimization. *IEEE Trans. Signal Inf. Process. over Netw.*, 6:774–789, 2020.

[28] K. Yuan, Q. Ling, and W. Yin. On the convergence of decentralized gradient descent. *SIAM J. Optim.*, 26(3):1835–1854, 2016.

[29] B. Sirb and X. Ye. Decentralized consensus algorithm with delayed and stochastic gradients. *SIAM J. Optim.*, 28(2):1232–1254, 2018.

[30] Z. Jiang, A. Balu, C. Hegde, and S. Sarkar. Collaborative deep learning in fixed topology networks. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 5904–5914, 2017.

[31] J. Wang and G. Joshi. Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms. In *Proc. ICML Workshop Coding Theory Mach. Learn.*, 2019.

[32] X. Li, W. Yang, S. Wang, and Z. Zhang. Communication-efficient local decentralized sgd methods. *arXiv preprint arXiv:1910.09126*, 2019.

[33] K. Tsianos, S. Lawlor, and M. Rabbat. Communication/computation tradeoffs in consensus-based distributed optimization. *Proc. Conf. Neural Inf. Process. Syst.*, 25:1943–1951, 2012.

[34] A. Koloskova, S. Stich, and M. Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *Proc. 36st Int. Conf. Mach. Learn.*, pages 3478–3487. PMLR, 2019.

[35] S. U. Stich. Local sgd converges fast and communicates little. In *Proc. ICLR*, 2018.

[36] P. Jiang and G. Agrawal. A linear speedup analysis of distributed deep learning with sparse and quantized communication. In *Proc. Adv. Neural Inf. Process. Syst.,*, pages 2525–2536, 2018.

[37] H. Yu, R. Jin, and S. Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed nonconvex optimization. In *Proc. 36st Int. Conf. Mach. Learn.*, pages 7184–7193. PMLR, 2019.

[38] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[39] L. Xiao and S. Boyd. Fast linear iterations for distributed averaging. In *Proc. 42nd IEEE Conf. Decision and Control*, volume 5, pages 4997–5002. IEEE, 2003.

[40] A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proc. 29th Int. Conf. Mach. Learn.*, pages 1571–1578, 2012.

[41] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

[42] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009.

[43] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge Univ. Press, 2012.

APPENDIX A
PROOF OF PROPOSITION 1

In order to prove Proposition 1, we first present the preliminaries of some notations for facilitating the proof, and then give a supporting lemma of DFL convergence where the error

convergence bound of the synchronous SGD and the local drift brought by DFL are presented. Lastly, we provide the proof of Proposition 1 based on the supporting lemma.

A. Preliminaries of Proof

For convenience, we first define Ξ_t to denote the set $\{\xi_t^{(1)}, \xi_t^{(2)}, \ldots, \xi_t^{(N)}\}$ of mini-batches at N nodes in iteration t. We use $\mathbf{E}_t$ to denote the conditional expectation $\mathbb{E}_{\Xi_t|\mathbf{X}_t}$. In addition, for $t \in [k]_1$, we define the averaged stochastic gradient and the averaged global gradient as the following:

$$\mathcal{G}_t = \frac{1}{N} \sum_{i=1}^{N} g(\mathbf{w}_t^{(i)}), \tag{33}$$

$$\mathcal{H}_t = \frac{1}{N} \sum_{i=1}^{N} \nabla F(\mathbf{w}_t^{(i)}). \tag{34}$$

And when $t \in [k]_2$, DFL performs inter-node communication without gradient descent so that $\mathcal{G}_t = 0, \mathcal{H}_t = 0$ for $t \in [k]_2$. Similar to the definition of $\mathbf{X}_t, \mathbf{G}_t$, we define the global gradient matrix $\nabla F(\mathbf{X}_t)$ as follows:

$$\nabla F(\mathbf{X}_t) \triangleq [\nabla F(\mathbf{w}_t^{(1)}), \nabla F(\mathbf{w}_t^{(2)}), \ldots, \nabla F(\mathbf{w}_t^{(N)})]. \tag{35}$$

We further define the Frobenius norm and the operator norm of matrix as follows to make analysis easier.

Definition 2 (The Frobenius norm [43]). The Frobenius norm defined for matrix $\mathbf{A} \in M_n$ by

$$\|\mathbf{A}\|_{\mathrm{F}}^2 = |\mathrm{Tr}(\mathbf{A}\mathbf{A}^\top)| = \sum_{i,j=1}^{n} |a_{ij}|^2. \tag{36}$$

Definition 3 (The operator norm [43]). The operator norm defined for matrix $\mathbf{A} \in M_n$ by

$$\|\mathbf{A}\|_{\mathrm{op}} = \max_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\| = \sqrt{\lambda_{\max}(\mathbf{A}^\top\mathbf{A})}. \tag{37}$$

According to the definition 2, the Frobenius norm of the global gradient matrix is

$$\|\nabla F(\mathbf{X}_t)\|_{\mathrm{F}}^2 = \sum_{i=1}^{N} \|\nabla F(\mathbf{w}_t^{(i)})\|^2.$$

B. The Supporting Lemma

We provide the supporting lemma for the proof of Proposition 1. The lemma establishes the basic convergence of DFL and the error convergence bound consists of two parts which are caused by the synchronous SGD and the local drift, respectively. We present the supporting lemma as follows:

Lemma 1. *Under Assumption 1, if the learning rate satisfies the condition $\eta L(1 + \frac{\beta}{N}) \le 1$ and all local models have*

the same initial point $\mathbf{w}_1$, then for Algorithm 1 of DFL, the average-squared gradient of T iterations is bounded as

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{u}_t)\|^2\right] \le \underbrace{\frac{2[F(\mathbf{w}_1) - F_{\mathrm{inf}}]}{\eta T} + \frac{\eta L\sigma^2}{N}}_{\text{synchronous SGD}} +$$

$$\underbrace{\frac{L^2}{T}\sum_{t=1}^{T}\frac{\mathbb{E}\|\mathbf{X}_t(\mathbf{I} - \mathbf{J})\|_{\mathrm{F}}^2}{N}}_{\text{local drift}}, \tag{38}$$

where $\mathbf{u}_t$ is defined in (17) and matrices $\mathbf{I}, \mathbf{J} \in \mathbb{R}^{N \times N}$.

In order to prove Lemma 1, we first present 3 supporting lemmas as follows. Then based on the presented lemmas, we finish the proof of Lemma 1.

Lemma 2. *Under the condition 4 and 5 of Assumption 1, we have the following upper bound for the expectation of the discrepancy between the averaged stochastic gradient $\mathcal{G}_t$ and the averaged global gradient $\mathcal{H}_t$ as*

$$\mathbb{E}_{\Xi_t|\mathbf{X}_t}[\|\mathcal{G}_t - \mathcal{H}_t\|^2] \le \frac{\beta}{N^2}\|\nabla F(\mathbf{X}_t)\|_{\mathrm{F}}^2 + \frac{\sigma^2}{N}. \tag{39}$$

Proof. According to the definition of $\mathcal{G}_t$ and $\mathcal{H}_t$ in (33) and (34), respectively, we have

$$\mathbb{E}_{\Xi_t|\mathbf{X}_t}[\|\mathcal{G}_t - \mathcal{H}_t\|^2]$$

$$= \mathbb{E}_{\Xi_t|\mathbf{X}_t}\left\|\frac{1}{N}\sum_{i=1}^{N}\left[g(\mathbf{w}_t^{(i)}) - \nabla F(\mathbf{w}_t^{(i)})\right]\right\|^2$$

$$= \frac{1}{N^2}\mathbb{E}_{\Xi_t|\mathbf{X}_t}\left[\sum_{i=1}^{N}\left\|g(\mathbf{w}_t^{(i)}) - \nabla F(\mathbf{w}_t^{(i)})\right\|^2\right] +$$

$$\frac{1}{N^2}\mathbb{E}_{\Xi_t|\mathbf{X}_t}\sum_{j\neq l}^{N}\left\langle g(\mathbf{w}_t^{(j)}) - \nabla F(\mathbf{w}_t^{(j)}), g(\mathbf{w}_t^{(l)}) - \nabla F(\mathbf{w}_t^{(l)})\right\rangle$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}_{\xi_t^{(i)}|\mathbf{X}_t}\left\|g(\mathbf{w}_t^{(i)}) - \nabla F(\mathbf{w}_t^{(i)})\right\|^2 +$$

$$\frac{1}{N^2}\sum_{j\neq l}^{N}\left\langle \mathbb{E}_{\xi_t^{(j)}|\mathbf{X}_t}\left[g(\mathbf{w}_t^{(j)}) - \nabla F(\mathbf{w}_t^{(j)})\right],\right.$$

$$\left. \mathbb{E}_{\xi_t^{(l)}|\mathbf{X}_t}\left[g(\mathbf{w}_t^{(l)}) - \nabla F(\mathbf{w}_t^{(l)})\right]\right\rangle,$$

where the last equation is due to $\{\xi_t^{(i)}\}$ are independent random variables. Then, according to the condition 4 in Assumption 1, we can obtain that the inner product term of all across terms in the last equation is zero. Thus, we have

$$\mathbb{E}_{\Xi_t|\mathbf{X}_t}[\|\mathcal{G}_t - \mathcal{H}_t\|^2] \le \frac{1}{N^2}\sum_{i=1}^{N}\left[\beta\left\|\nabla F(\mathbf{w}_t^{(i)})\right\|^2 + \sigma^2\right]$$

$$= \frac{\beta}{N}\frac{\|\nabla F(\mathbf{X}_t)\|_{\mathrm{F}}^2}{N} + \frac{\sigma^2}{N},$$

where the last equation is due to the condition 5 in Assumption 1 and the definition of the Frobenius norm of $\nabla F(\mathbf{X}_t)$. $\square$

Lemma 3. *Based on Assumption 1, the expectation of the inner product between stochastic gradient and global gradient can expanded as*

$$\mathbf{E}_t[\langle \nabla F(\mathbf{u}_t), \mathcal{G}_t \rangle] = \frac{1}{2}\|\nabla F(\mathbf{u}_t)\|^2 + \frac{1}{2N}\sum_{i=1}^{N}\left\|\nabla F(\mathbf{w}_t^{(i)})\right\|^2$$
$$- \frac{1}{2N}\sum_{i=1}^{N}\left\|\nabla F(\mathbf{u}_t) - \nabla F(\mathbf{w}_t^{(i)})\right\|^2, \tag{40}$$

where $\mathbf{E}_t$ is the conditional expectation $\mathbb{E}_{\Xi_t|\mathbf{X}_t}$.

Proof. We can directly prove equation (40) as follows:

$$\mathbf{E}_t[\langle \nabla F(\mathbf{u}_t), \mathcal{G}_t \rangle]$$
$$= \mathbf{E}_t\left[\left\langle \nabla F(\mathbf{u}_t), \frac{1}{N}\sum_{i=1}^{N} g(\mathbf{w}_t^{(i)}) \right\rangle\right]$$
$$= \frac{1}{N}\sum_{i=1}^{N}\left\langle \nabla F(\mathbf{u}_t), \nabla F(\mathbf{w}_t^{(i)}) \right\rangle$$
$$= \frac{1}{2N}\sum_{i=1}^{N}\left[\|\nabla F(\mathbf{u}_t)\|^2 + \left\|\nabla F(\mathbf{w}_t^{(i)})\right\|^2 - \left\|\nabla F(\mathbf{u}_t) - \nabla F(\mathbf{w}_t^{(i)})\right\|^2\right] \tag{41}$$

$$= \frac{1}{2}\|\nabla F(\mathbf{u}_t)\|^2 + \frac{1}{2N}\sum_{i=1}^{N}\left\|\nabla F(\mathbf{w}_t^{(i)})\right\|^2$$
$$- \frac{1}{2N}\sum_{i=1}^{N}\left\|\nabla F(\mathbf{u}_t) - \nabla F(\mathbf{w}_t^{(i)})\right\|^2, \tag{42}$$

where equation (41) is because $\mathbf{a}^\top \mathbf{b} = \frac{1}{2}(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a} - \mathbf{b}\|^2)$. □

Lemma 4. *Based on the Assumption 1, the squared norm of stochastic gradient can be bounded as*

$$\mathbf{E}_t\left[\|\mathcal{G}_t\|^2\right] \leq \left(\frac{\beta}{N} + 1\right)\frac{\|\nabla F(\mathbf{X}_t)\|_{\mathrm{F}}^2}{N} + \frac{\sigma^2}{N}. \tag{43}$$

Proof. According to the condition 4 of Assumption 1, we have $\mathbf{E}_t[\mathcal{G}_t] = \mathcal{H}_t$. Then we obtain

$$\mathbf{E}_t\left[\|\mathcal{G}_t\|^2\right] = \mathbf{E}_t\left[\|\mathcal{G}_t - \mathbf{E}_t[\mathcal{G}_t]\|^2\right] + \|\mathbf{E}_t[\mathcal{G}_t]\|^2 \tag{44}$$
$$= \mathbf{E}_t\left[\|\mathcal{G}_t - \mathcal{H}_t\|^2\right] + \|\mathcal{H}_t\|^2 \tag{45}$$
$$\leq \frac{\beta}{N}\frac{\|\nabla F(\mathbf{X}_t)\|_{\mathrm{F}}^2}{N} + \frac{\sigma^2}{N} + \|\mathcal{H}_t\|^2 \tag{46}$$
$$\leq \frac{\beta}{N}\frac{\|\nabla F(\mathbf{X}_t)\|_{\mathrm{F}}^2}{N} + \frac{\sigma^2}{N} + \frac{1}{N}\|\nabla F(\mathbf{X}_t)\|_{\mathrm{F}}^2 \tag{47}$$
$$= \left(\frac{\beta}{N} + 1\right)\frac{\|\nabla F(\mathbf{X}_t)\|_{\mathrm{F}}^2}{N} + \frac{\sigma^2}{N}, \tag{48}$$

where (46) is from Lemma 2, and (47) is because the convexity of vector norm and Jensen's inequality:

$$\|\mathcal{H}_t\|^2 = \left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F(\mathbf{w}_t^{(i)})\right\|^2$$
$$\leq \frac{1}{N}\sum_{i=1}^{N}\left\|\nabla F(\mathbf{w}_t^{(i)})\right\|^2$$
$$= \frac{1}{N}\|\nabla F(\mathbf{X}_t)\|_{\mathrm{F}}^2.$$

□

Obviously, Lemmas 2, 3, 4 are established under the condition $t \in [k]_1$. However, Lemmas 2, 3, 4 still hold under the communication stage for $t \in [k]_2$. We argue that as follows in detail. For the equation (39) of Lemma 2, when $t \in [k]_2$, we can get the left side of the inequality $\mathbb{E}_{\Xi_t|\mathbf{X}_t}[\|\mathcal{G}_t - \mathcal{H}_t\|^2] = 0$, and the right side of the inequality is greater than 0. Thus, Lemma 2 is true for $t \in [k]_2$. Considering equation (40) of Lemma 3, for $t \in [k]_2$, we get the left side of the equality $\mathbf{E}_t[\langle \nabla F(\mathbf{u}_t), \mathcal{G}_t \rangle] = 0$ and the right side is $\frac{1}{2}\|\nabla F(\mathbf{u}_t)\|^2 - \frac{1}{2N}\sum_{i=1}^{N}\|\nabla F(\mathbf{u}_t)\|^2 = 0$ due to there exists no gradient at inter-node communication stage. Therefore, Lemma 3 is established for $t \in [k]_2$. Focusing on Lemma 4, for $t \in [k]_2$, we can get the left side of (43) $\mathbf{E}_t\left[\|\mathcal{G}_t\|^2\right] = 0$, and the right side $\left(\frac{\beta}{N} + 1\right)\frac{\|\nabla F(\mathbf{X}_t)\|_{\mathrm{F}}^2}{N} + \frac{\sigma^2}{N} > 0$. Thus for $t \in [k]_2$, Lemma 4 holds.

After presenting the three lemmas, we begin to prove Lemma 1.

Proof of Lemma 1: According to the L-smooth assumption from Assumption 1, we have

$$\mathbf{E}_t[F(\mathbf{u}_{t+1})] - F(\mathbf{u}_t) \leq -\eta\mathbf{E}_t[\langle \nabla F(\mathbf{u}_t), \mathcal{G}_t \rangle] + \frac{\eta^2 L}{2}\mathbf{E}_t\left[\|\mathcal{G}_t\|^2\right]. \tag{49}$$

Note that $\mathbf{u}_{t+1} = \mathbf{u}_t$ for $t \in [k]_2$ and $\mathbf{E}_t[F(\mathbf{u}_{t+1})] - F(\mathbf{u}_t) = 0$. Because $\mathcal{G}_t = 0$ for $t \in [k]_2$, the right side of (49) is 0. Thus (49) holds for $t \in [k]_2$. Combining with Lemma 3 and 4, one can get

$$\mathbf{E}_t[F(\mathbf{u}_{t+1})] - F(\mathbf{u}_t)$$
$$\leq -\frac{\eta}{2}\|\nabla F(\mathbf{u}_t)\|^2 - \frac{\eta}{2N}\sum_{i=1}^{N}\left\|\nabla F(\mathbf{w}_t^{(i)})\right\|^2 +$$
$$\frac{\eta}{2N}\sum_{i=1}^{N}\left\|\nabla F(\mathbf{u}_t) - \nabla F(\mathbf{w}_t^{(i)})\right\|^2 +$$
$$\frac{\eta^2 L}{2N}\sum_{i=1}^{N}\left\|\nabla F(\mathbf{w}_t^{(i)})\right\|^2 \cdot \left(\frac{\beta}{N} + 1\right) + \frac{\eta^2 L\sigma^2}{2N}$$
$$\leq -\frac{\eta}{2}\|\nabla F(\mathbf{u}_t)\|^2 - \frac{\eta}{2}\left[1 - \eta L\left(\frac{\beta}{N} + 1\right)\right]\cdot\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla F(\mathbf{w}_t^{(i)})\right\|^2$$
$$+ \frac{\eta^2 L\sigma^2}{2N} + \frac{\eta L^2}{2N}\sum_{i=1}^{N}\left\|\mathbf{u}_t - \mathbf{w}_t^{(i)}\right\|^2.$$

By rearranging the above inequality and according to the definition of Frobenius norm, we further get

$$\|\nabla F(\mathbf{u}_t)\|^2 \leq \frac{2[F(\mathbf{u}_t) - \mathbf{E}_t[F(\mathbf{u}_{t+1})]]}{\eta} + \frac{\eta L\sigma^2}{N} +$$
$$\frac{L^2}{N}\sum_{i=1}^{N}\left\|\mathbf{u}_t - \mathbf{w}_t^{(i)}\right\|^2 - \left[1 - \eta L\left(\frac{\beta}{N} + 1\right)\right]\frac{1}{N}\|\nabla F(\mathbf{X}_t)\|_{\mathrm{F}}^2. \tag{50}$$

Taking the total expectation and averaging over all iterations, we can get

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{u}_t)\|^2\right] \leq \frac{2[F(\mathbf{u}_1) - F_{\inf}]}{\eta T} + \frac{\eta L \sigma^2}{N}$$
$$+ \frac{L^2}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\left\|\mathbf{u}_t - \mathbf{w}_t^{(i)}\right\|^2$$
$$- \frac{1}{NT}\left[1 - \eta L\left(\frac{\beta}{N} + 1\right)\right]\sum_{t=1}^{T}\mathbb{E}\|\nabla F(\mathbf{X}_t)\|_{\mathrm{F}}^2. \quad (51)$$

Recalling the definition $\mathbf{u}_t = \mathbf{X}_t\frac{1}{N}$, we can obtain

$$\sum_{i=1}^{N}\left\|\mathbf{u}_t - \mathbf{w}_t^{(i)}\right\|^2 = \left\|\mathbf{u}_t\mathbf{1}^\top - \mathbf{X}_t\right\|_{\mathrm{F}}^2$$
$$= \left\|\mathbf{X}_t\frac{\mathbf{1}\mathbf{1}^\top}{N} - \mathbf{X}_t\right\|_{\mathrm{F}}^2$$
$$= \|\mathbf{X}_t(\mathbf{I} - \mathbf{J})\|_{\mathrm{F}}^2. \quad (52)$$

By substituting (52) into (51) and according to the condition $\eta L\left(1 + \frac{\beta}{N}\right) \leq 1$, we complete the proof of Lemma 1.

C. Proof of Proposition 1

We first provide three lemmas for assisting proof as follows.

Lemma 5. *For two real matrices $\mathbf{A} \in \mathbb{R}^{d \times m}$ and $\mathbf{B}^{m \times m}$, if $\mathbf{B}$ is symmetric, then we have*

$$\|\mathbf{AB}\|_{\mathrm{F}} \leq \|\mathbf{B}\|_{\mathrm{op}}\|\mathbf{A}\|_{\mathrm{F}}. \quad (53)$$

Proof. Assuming that $\mathbf{a}_1^\top, \ldots, \mathbf{a}_d^\top$ denote the rows of matrix $\mathbf{A}$ and $\mathcal{I} = \{i \in [1, d] : \|\mathbf{a}_i\| \neq 0\}$. Then we obtain

$$\|\mathbf{AB}\|_{\mathrm{F}}^2 = \sum_{i=1}^{d}\|\mathbf{a}_i^\top\mathbf{B}\|^2 = \sum_{i\in\mathcal{I}}^{d}\|\mathbf{Ba}_i\|^2$$
$$= \sum_{i\in\mathcal{I}}^{d}\frac{\|\mathbf{Ba}_i\|^2}{\|\mathbf{a}_i\|^2}\|\mathbf{a}_i\|^2$$
$$\leq \sum_{i\in\mathcal{I}}^{d}\|\mathbf{B}\|_{\mathrm{op}}^2\|\mathbf{a}_i\|^2 = \|\mathbf{B}\|_{\mathrm{op}}^2\sum_{i\in\mathcal{I}}^{d}\|\mathbf{a}_i\|^2 = \|\mathbf{B}\|_{\mathrm{op}}^2\|\mathbf{A}\|_{\mathrm{F}}^2,$$

where the last inequality is due to Definition 3 about matrix operator norm. $\square$

Lemma 6. *Consider two matrices $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times m}$. We have*

$$|\mathrm{Tr}(\mathbf{AB})| \leq \|\mathbf{A}\|_{\mathrm{F}}\|\mathbf{B}\|_{\mathrm{F}}. \quad (54)$$

Proof. Assume $\mathbf{a}_i^\top \in \mathbb{R}^n$ is the i-th row of matrix $\mathbf{A}$ and $\mathbf{b}_i^\top \in \mathbb{R}^n$ is the i-th column of matrix $\mathbf{B}$. According to the definition of matrix trace, we have

$$\mathrm{Tr}(\mathbf{AB}) = \sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{A}_{ij}\mathbf{B}_{ji}$$
$$= \sum_{i=1}^{m}\mathbf{a}_i^\top\mathbf{b}_i. \quad (55)$$

Then according to Cauchy-Schwartz inequality, we further obtain

$$|\sum_{i=1}^{m}\mathbf{a}_i^\top\mathbf{b}_i|^2 \leq \left(\sum_{i=1}^{m}\|\mathbf{a}_i\|^2\right)\left(\sum_{i=1}^{m}\|\mathbf{b}_i\|^2\right)$$
$$= \|\mathbf{A}\|_{\mathrm{F}}^2\|\mathbf{B}\|_{\mathrm{F}}^2. \quad (56)$$

Then combining (55) and (56), we finish the proof. $\square$

Lemma 7. *Consider a matrix $\mathbf{C} \in \mathbb{R}^{m \times m}$ which satisfies condition 6 of Assumption 1. Then we have*

$$\|\mathbf{C}^j - \mathbf{J}\|_{\mathrm{op}} = \zeta^j, \quad (57)$$

where $\zeta = \max\{|\lambda_2(\mathbf{C})|, |\lambda_m|(\mathbf{C})\}$.

Proof. Because $\mathbf{C}$ is a real symmetric matrix, it can be decomposed as $\mathbf{C} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$, where $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1(\mathbf{C}), \lambda_2(\mathbf{C}), \ldots, \lambda_m(\mathbf{C})\}$. Similarly, matrix $\mathbf{J}$ can be decomposed as $\mathbf{J} = \mathbf{Q}\mathbf{\Lambda}_0\mathbf{Q}^\top$ where $\mathbf{\Lambda}_0 = \mathrm{diag}\{1, 0, \ldots, 0\}$. Then we have

$$\mathbf{C}^j - \mathbf{J} = (\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top)^j - \mathbf{J} = \mathbf{Q}(\mathbf{\Lambda}^j - \mathbf{\Lambda}_0)\mathbf{Q}^\top. \quad (58)$$

According to the definition of the matrix operator norm, we further obtain

$$\|\mathbf{C}^j - \mathbf{J}\|_{\mathrm{op}} = \sqrt{\lambda_{\max}((\mathbf{C}^j - \mathbf{J})^\top(\mathbf{C}^j - \mathbf{J}))}$$
$$= \sqrt{\lambda_{\max}(\mathbf{C}^{2j} - \mathbf{J})},$$

where the last equality is due to

$$(\mathbf{C}^j - \mathbf{J})^\top(\mathbf{C}^j - \mathbf{J}) = (\mathbf{C}^j - \mathbf{J})(\mathbf{C}^j - \mathbf{J})$$
$$= \mathbf{C}^{2j} + \mathbf{J}^2 - \mathbf{C}^j\mathbf{J} - \mathbf{J}\mathbf{C}^j$$
$$= \mathbf{C}^{2j} + \mathbf{J} - 2\mathbf{J} = \mathbf{C}^{2j} - \mathbf{J}.$$

Because $\mathbf{C}^{2j} - \mathbf{J} = \mathbf{Q}(\mathbf{\Lambda}^{2j} - \mathbf{\Lambda}_0)\mathbf{Q}^\top$ from (58), the maximum eigenvalue is $\max\{0, \lambda_2(\mathbf{C})^{2j}, \ldots, \lambda_m(\mathbf{C})^{2j}\} = \zeta^{2j}$ where ζ is the second largest eigenvalue of $\mathbf{C}$. Therefore, we have

$$\|\mathbf{C}^j - \mathbf{J}\|_{\mathrm{op}} = \sqrt{\lambda_{\max}(\mathbf{C}^{2j} - \mathbf{J})} = \zeta^j.$$
$\square$

Proof of Proposition 1: After introducing the three lemmas, we begin to prove Proposition 1. We recall the result of Lemma 1 as follows:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{u}_t)\|^2\right] \leq \frac{2[F(\mathbf{w}_1) - F_{\inf}]}{\eta T} + \frac{\eta L \sigma^2}{N} +$$
$$\frac{L^2}{TN}\sum_{t=1}^{T}\mathbb{E}\|\mathbf{X}_t(\mathbf{I} - \mathbf{J})\|_{\mathrm{F}}^2. \quad (59)$$

Our target is to find a upper bound of the local drift term $\frac{L^2}{TN}\sum_{t=1}^{T}\mathbb{E}\|\mathbf{X}_t(\mathbf{I} - \mathbf{J})\|_{\mathrm{F}}^2$. First of all, we consider the term $\mathbb{E}\|\mathbf{X}_t(\mathbf{I} - \mathbf{J})\|_{\mathrm{F}}^2$.

According to the learning strategy of DFL in (5), we have

$$\mathbf{X}_t(\mathbf{I} - \mathbf{J}) = (\mathbf{X}_{t-1} - \eta\mathbf{G}_{t-1}')\mathbf{C}_{t-1}(\mathbf{I} - \mathbf{J})$$
$$= \mathbf{X}_{t-1}(\mathbf{I} - \mathbf{J})\mathbf{C}_{t-1} - \eta\mathbf{G}_{t-1}'(\mathbf{C}_{t-1} - \mathbf{J}),$$

where the last equality is due to $\mathbf{C}_{t-1}\mathbf{J} = \mathbf{J}\mathbf{C}_{t-1} = \mathbf{J}$ and hence $(\mathbf{I} - \mathbf{J})\mathbf{C}_{t-1} = \mathbf{C}_{t-1}(\mathbf{I} - \mathbf{J})$. We further expand the expression of $\mathbf{X}_{t-1}$ as follows:

$$
\begin{aligned}
\mathbf{X}_t(\mathbf{I} - \mathbf{J}) =&[\mathbf{X}_{t-2}(\mathbf{I} - \mathbf{J})\mathbf{C}_{t-2} - \eta\mathbf{G}'_{t-2}(\mathbf{C}_{t-2} - \mathbf{J})]\mathbf{C}_{t-1}\\
&- \eta\mathbf{G}'_{t-1}(\mathbf{C}_{t-1} - \mathbf{J})\\
=&\mathbf{X}_{t-2}(\mathbf{I} - \mathbf{J})\mathbf{C}_{t-2}\mathbf{C}_{t-1} - \eta\mathbf{G}'_{t-2}(\mathbf{C}_{t-2}\mathbf{C}_{t-1} - \mathbf{J})\\
&- \eta\mathbf{G}'_{t-1}(\mathbf{C}_{t-1} - \mathbf{J}).
\end{aligned}
$$

Repeating the same expansion for $\mathbf{X}_{t-2}, \mathbf{X}_{t-3}, \ldots, \mathbf{X}_2$, finally we obtain

$$
\mathbf{X}_t(\mathbf{I} - \mathbf{J}) = \mathbf{X}_1(\mathbf{I} - \mathbf{J})\boldsymbol{\Phi}_{1,t-1} - \eta\sum_{s=1}^{t-1}\mathbf{G}'_s(\boldsymbol{\Phi}_{s,t-1} - \mathbf{J}),
\tag{60}
$$

where $\boldsymbol{\Phi}_{s,t-1} = \prod_{l=s}^{t-1}\mathbf{C}_l$. As parameters of all local models are initialized at the same point $\mathbf{X}_1(\mathbf{I} - \mathbf{J}) = 0$, the squared norm of the local drift term can be rewritten as

$$
\mathbb{E}\|\mathbf{X}_t(\mathbf{I} - \mathbf{J})\|_\mathrm{F}^2 = \eta^2\mathbb{E}\left\|\sum_{s=1}^{t-1}\mathbf{G}'_s(\boldsymbol{\Phi}_{s,t-1} - \mathbf{J})\right\|_\mathrm{F}^2.
\tag{61}
$$

Then we list the expression of $\boldsymbol{\Phi}_{s,t-1}$ for any $s \leq t - 1$. Without loss of generality, we assume $t = j\tau + i$ where $i \leq \tau_1$ according to Algorithm 1. Note that j denotes the index of iteration round and i denote the index of local updates. The expression of matrix $\boldsymbol{\Phi}_{s,t-1}$ is presented as follows:

$$
\boldsymbol{\Phi}_{s,t-1} =
\begin{cases}
\mathbf{I}, & j\tau < s \leq j\tau + i\\
\mathbf{C}^{j\tau-s}, & (j-1)\tau + \tau_1 < s \leq j\tau\\
\mathbf{C}^{\tau_2}, & (j-1)\tau < s \leq (j-1)\tau + \tau_1\\
\mathbf{C}^{\tau_2}\cdot\mathbf{C}^{(j-1)\tau-s}, & (j-2)\tau + \tau_1 < s \leq (j-1)\tau\\
\mathbf{C}^{2\tau_2}, & (j-2)\tau < s \leq (j-2)\tau + \tau_1\\
\vdots &\\
\mathbf{C}^{(j-1)\tau_2}\cdot\mathbf{C}^{\tau-s}, & \tau_1 < s \leq \tau\\
\mathbf{C}^{j\tau_2}, & 0 < s \leq \tau_1.
\end{cases}
\tag{62}
$$

For the simplification of writing, we define the sum of stochastic gradient within one round as $\mathbf{Y}_r = \sum_{s=r\tau+1}^{r\tau+\tau_1}\mathbf{G}'_s = \sum_{s=r\tau+1}^{r\tau+\tau_1}\mathbf{G}_s$ for $0 \leq r < j$ and $\mathbf{Y}_j = \sum_{s=j\tau+1}^{j\tau+i-1}\mathbf{G}_s$. Similarly, we define accumulated global gradient $\mathbf{Q}_r = \sum_{s=r\tau+1}^{r\tau+\tau_1}\nabla F(\mathbf{X}_s)$ for $0 \leq r < j$ and $\mathbf{Q}_j = \sum_{s=j\tau+1}^{j\tau+i-1}\nabla F(\mathbf{X}_s)$. Thus, we have

$$
\sum_{s=1}^{\tau}\mathbf{G}'_s(\boldsymbol{\Phi}_{s,t-1} - \mathbf{J}) = \sum_{s=1}^{\tau_1}\mathbf{G}_s(\mathbf{C}^{j\tau_2} - \mathbf{J}) = \mathbf{Y}_0(\mathbf{C}^{j\tau_2} - \mathbf{J})
$$

$$
\sum_{s=\tau+1}^{2\tau}\mathbf{G}'_s(\boldsymbol{\Phi}_{s,t-1} - \mathbf{J}) = \mathbf{Y}_1(\mathbf{C}^{(j-1)\tau_2} - \mathbf{J})
$$

$$
\vdots
$$

$$
\sum_{s=j\tau+1}^{j\tau+i-1}\mathbf{G}'_s(\boldsymbol{\Phi}_{s,t-1} - \mathbf{J}) = \mathbf{Y}_j(\mathbf{I} - \mathbf{J}).
$$

Hence, we sum all these terms and obtain

$$
\sum_{s=1}^{t-1}\mathbf{G}'_s(\boldsymbol{\Phi}_{s,t-1} - \mathbf{J}) = \sum_{r=0}^{j}\mathbf{Y}_r(\mathbf{C}^{(j-r)\tau_2} - \mathbf{J}).
\tag{63}
$$

We further decompose the local drift term into two parts as follows:

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{X}_t(\mathbf{I} - \mathbf{J})\|_\mathrm{F}^2\\
=&\eta^2\mathbb{E}\left\|\sum_{r=0}^{j}\mathbf{Y}_r(\mathbf{C}^{(j-r)\tau_2} - \mathbf{J})\right\|_\mathrm{F}^2\\
=&\eta^2\mathbb{E}\left\|\sum_{r=0}^{j}(\mathbf{Y}_r - \mathbf{Q}_r)(\mathbf{C}^{(j-r)\tau_2} - \mathbf{J}) + \sum_{r=0}^{j}\mathbf{Q}_r(\mathbf{C}^{(j-r)\tau_2} - \mathbf{J})\right\|_\mathrm{F}^2\\
\leq&2\eta^2\mathbb{E}\underbrace{\left\|\sum_{r=0}^{j}(\mathbf{Y}_r - \mathbf{Q}_r)(\mathbf{C}^{(j-r)\tau_2} - \mathbf{J})\right\|_\mathrm{F}^2}_{A_1}\\
&+ 2\eta^2\mathbb{E}\underbrace{\left\|\sum_{r=0}^{j}\mathbf{Q}_r(\mathbf{C}^{(j-r)\tau_2} - \mathbf{J})\right\|_\mathrm{F}^2}_{A_2},
\end{aligned}
\tag{64}
$$

where (64) is due to $\|\mathbf{a} + \mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$. Based on (64), we aim to separately derive the upper bounds for A_1 and A_2 before completing the proof of Proposition 1.

The Bound of A_1: For the term A_1 in (64), we have

$$
\begin{aligned}
&A_1\\
=&2\eta^2\mathbb{E}\left\|\sum_{r=0}^{j}(\mathbf{Y}_r - \mathbf{Q}_r)(\mathbf{C}^{(j-r)\tau_2} - \mathbf{J})\right\|_\mathrm{F}^2\\
=&2\eta^2\sum_{r=0}^{j}\mathbb{E}\|(\mathbf{Y}_r - \mathbf{Q}_r)(\mathbf{C}^{(j-r)\tau_2} - \mathbf{J})\|_\mathrm{F}^2\\
&+2\eta^2\sum_{n=0}^{j}\sum_{\substack{l=0\\l\neq n}}^{j}\mathbb{E}[\mathrm{Tr}((\mathbf{C}^{(j-n)\tau_2} - \mathbf{J})(\mathbf{Y}_n - \mathbf{Q}_n)^\top(\mathbf{Y}_l - \mathbf{Q}_l)(\mathbf{C}^{(j-l)\tau_2} - \mathbf{J}))]\\
=&2\eta^2\sum_{r=0}^{j}\mathbb{E}\|(\mathbf{Y}_r - \mathbf{Q}_r)(\mathbf{C}^{(j-r)\tau_2} - \mathbf{J})\|_\mathrm{F}^2
\end{aligned}
\tag{65}
$$

$$
\leq 2\eta^2\sum_{r=0}^{j}\mathbb{E}\|\mathbf{Y}_r - \mathbf{Q}_r\|_\mathrm{F}^2\|\mathbf{C}^{(j-r)\tau_2} - \mathbf{J}\|_\mathrm{op}^2
\tag{66}
$$

$$
= 2\eta^2\sum_{r=0}^{j}\mathbb{E}\|\mathbf{Y}_r - \mathbf{Q}_r\|_\mathrm{F}^2\zeta^{2\tau_2(j-r)}
\tag{67}
$$

$$
= 2\eta^2\sum_{r=0}^{j-1}\mathbb{E}\|\mathbf{Y}_r - \mathbf{Q}_r\|_\mathrm{F}^2\zeta^{2\tau_2(j-r)} + 2\eta^2\mathbb{E}\|\mathbf{Y}_j - \mathbf{Q}_j\|_\mathrm{F}^2,
\tag{68}
$$

where (66) comes from Lemma 6 and (67) is due to Lemma 7. Consider the inner product of the cross terms $g(\mathbf{w}_s^{(i)}) - \nabla F(\mathbf{w}_s^{(i)})$ and $g(\mathbf{w}_l^{(j)}) - \nabla F(\mathbf{w}_l^{(j)})$ for any $s \neq l$ and any i, j,

we can obtain

$$\mathbb{E}\left[\left\langle g(\mathbf{w}_s^{(i)}) - \nabla F(\mathbf{w}_s^{(i)}), g(\mathbf{w}_l^{(j)}) - \nabla F(\mathbf{w}_l^{(j)})\right\rangle\right]$$
$$=\mathbb{E}_{\mathbf{w}_s^{(i)}\xi_s^{(i)},\mathbf{w}_l^{(j)}}\mathbb{E}_{\xi_l^{(j)}|\mathbf{w}_s^{(i)}\xi_s^{(i)},\mathbf{w}_l^{(j)}}\left[\left\langle g(\mathbf{w}_s^{(i)})-\nabla F(\mathbf{w}_s^{(i)}), g(\mathbf{w}_l^{(j)})-\nabla F(\mathbf{w}_l^{(j)})\right\rangle\right]$$
$$=\mathbb{E}\left[\left\langle g(\mathbf{w}_s^{(i)})-\nabla F(\mathbf{w}_s^{(i)}), \mathbb{E}_{\xi_l^{(j)}|\mathbf{w}_s^{(i)}\xi_s^{(i)},\mathbf{w}_l^{(j)}}\left[g(\mathbf{w}_l^{(j)})-\nabla F(\mathbf{w}_l^{(j)})\right]\right\rangle\right]$$
$$=\mathbb{E}\left[\left\langle g(\mathbf{w}_s^{(i)})-\nabla F(\mathbf{w}_s^{(i)}), 0\right\rangle\right] \tag{69}$$
$$=0, \tag{70}$$

where (69) comes from condition 4 of Assumption 1. Therefore, according to (70), we can easily obtain $\mathbb{E}[(\mathbf{Y}_n - \mathbf{Q}_n)^\top(\mathbf{Y}_l - \mathbf{Q}_l)] = 0$. Thus (65) holds.

Following the result of (68), for and $0 \le r < j$, we have

$$\mathbb{E}\left[\|\mathbf{Y}_r - \mathbf{Q}_r\|_F^2\right]$$
$$=\mathbb{E}\left[\left\|\sum_{s=r\tau+1}^{r\tau+\tau_1}[\mathbf{G}_s - \nabla F(\mathbf{X}_s)]\right\|_F^2\right] \tag{71}$$
$$=\sum_{i=1}^N\mathbb{E}\left[\left\|\sum_{s=r\tau+1}^{r\tau+\tau_1}[g(\mathbf{w}_s^{(i)}) - \nabla F(\mathbf{w}_s^{(i)})]\right\|^2\right] \tag{72}$$
$$=\sum_{i=1}^N\mathbb{E}\left[\sum_{s=r\tau+1}^{r\tau+\tau_1}\left\|g(\mathbf{w}_s^{(i)}) - \nabla F(\mathbf{w}_s^{(i)})\right\|^2 + \right.$$
$$\left.\sum_{s\ne l}\left\langle g(\mathbf{w}_s^{(i)}) - \nabla F(\mathbf{w}_s^{(i)}), g(\mathbf{w}_l^{(i)}) - \nabla F(\mathbf{w}_l^{(i)})\right\rangle\right] \tag{73}$$
$$=\sum_{i=1}^N\mathbb{E}\left[\sum_{s=r\tau+1}^{r\tau+\tau_1}\left\|g(\mathbf{w}_s^{(i)}) - \nabla F(\mathbf{w}_s^{(i)})\right\|^2\right] \tag{74}$$
$$=\mathbb{E}\left[\sum_{s=r\tau+1}^{r\tau+\tau_1}\sum_{i=1}^N\left\|g(\mathbf{w}_s^{(i)}) - \nabla F(\mathbf{w}_s^{(i)})\right\|^2\right] \tag{75}$$
$$\le\beta\sum_{s=r\tau+1}^{r\tau+\tau_1}\sum_{i=1}^N\mathbb{E}\left[\left\|\nabla F(\mathbf{w}_s^{(i)})\right\|^2\right] + \tau_1 N\sigma^2 \tag{76}$$
$$=\beta\sum_{s=r\tau+1}^{r\tau+\tau_1}\|\nabla F(\mathbf{X}_s)\|_F^2 + \tau_1 N\sigma^2, \tag{77}$$

where (74) comes from (70) and (76) is according to condition 5 of Assumption 1. Similarly, we can obtain

$$\mathbb{E}\left[\|\mathbf{Y}_j - \mathbf{Q}_j\|_F^2\right] \le \beta\sum_{s=j\tau+1}^{j\tau+i-1}\|\nabla F(\mathbf{X}_s)\|_F^2 + (i-1)N\sigma^2. \tag{78}$$

Substituting (77) and (78) into (68), we obtain

$$A_1 \le 2\eta^2\sum_{r=0}^{j-1}\left[\zeta^{2\tau_2(j-r)}\left(\beta\sum_{s=r\tau+1}^{r\tau+\tau_1}\|\nabla F(\mathbf{X}_s)\|_F^2 + \tau_1 N\sigma^2\right)\right]$$
$$+ 2\eta^2\beta\sum_{s=j\tau+1}^{j\tau+i-1}\|\nabla F(\mathbf{X}_s)\|_F^2 + 2\eta^2(i-1)N\sigma^2 \tag{79}$$
$$\le 2\eta^2 N\sigma^2\left[\frac{\tau_1\zeta^{2\tau_2}}{1-\zeta^{2\tau_2}} + i - 1\right] +$$
$$2\eta^2\beta\sum_{r=0}^{j-1}\left[\zeta^{2\tau_2(j-r)}\left(\sum_{s=r\tau+1}^{r\tau+\tau_1}\|\nabla F(\mathbf{X}_s)\|_F^2\right)\right] +$$
$$2\eta^2\beta\sum_{s=j\tau+1}^{j\tau+i-1}\|\nabla F(\mathbf{X}_s)\|_F^2 \tag{80}$$
$$\le 2\eta^2 N\sigma^2\left[\frac{\tau_1\zeta^{2\tau_2}}{1-\zeta^{2\tau_2}} + \tau_1 - 1\right] +$$
$$2\eta^2\beta\sum_{r=0}^{j-1}\left[\zeta^{2\tau_2(j-r)}\left(\sum_{s=r\tau+1}^{r\tau+\tau_1}\|\nabla F(\mathbf{X}_s)\|_F^2\right)\right] +$$
$$2\eta^2\beta\sum_{s=j\tau+1}^{j\tau+\tau_1}\|\nabla F(\mathbf{X}_s)\|_F^2, \tag{81}$$

where (80) is because

$$\sum_{r=0}^{j-1}\zeta^{2\tau_2(j-r)} = \frac{\zeta^{2\tau_2}(1 - (\zeta^{2\tau_2})^j)}{1 - \zeta^{2\tau_2}} \le \frac{\zeta^{2\tau_2}}{1 - \zeta^{2\tau_2}},$$

and (81) comes from $i \le \tau_1$. Next, we sum over all iterations in one round from $i = 1$ to $i = \tau$ and obtain

$$\sum_{i=1}^\tau A_1 \le 2\eta^2 N\sigma^2\tau\left[\frac{\tau_1\zeta^{2\tau_2}}{1-\zeta^{2\tau_2}} + \tau_1 - 1\right]$$
$$+ 2\eta^2\beta\tau\sum_{r=0}^{j-1}\left[\zeta^{2\tau_2(j-r)}\left(\sum_{s=r\tau+1}^{r\tau+\tau_1}\|\nabla F(\mathbf{X}_s)\|_F^2\right)\right]$$
$$+ 2\eta^2\beta\tau\sum_{s=j\tau+1}^{j\tau+\tau_1}\|\nabla F(\mathbf{X}_s)\|_F^2 \tag{82}$$
$$= 2\eta^2 N\sigma^2\tau\left[\frac{\tau_1\zeta^{2\tau_2}}{1-\zeta^{2\tau_2}} + \tau_1 - 1\right]$$
$$+ 2\eta^2\beta\tau\sum_{r=0}^j\left[\zeta^{2\tau_2(j-r)}\left(\sum_{s=r\tau+1}^{r\tau+\tau_1}\|\nabla F(\mathbf{X}_s)\|_F^2\right)\right]. \tag{83}$$

Then we sum over all rounds from $j = 0$ to $j = \frac{T}{\tau} - 1$, where

T is the total iterations and we get

$$\sum_{j=0}^{\frac{T}{\tau}-1} \sum_{i=1}^{\tau} A_1$$

$$\leq 2\eta^2 N\sigma^2 T \left[\frac{\tau_1 \zeta^{2\tau_2}}{1-\zeta^{2\tau_2}} + \tau_1 - 1 \right] +$$

$$2\eta^2 \beta\tau \sum_{j=0}^{\frac{T}{\tau}-1} \sum_{r=0}^{j} \left[\zeta^{2\tau_2(j-r)} \left(\sum_{s=r\tau+1}^{r\tau+\tau_1} \|\nabla F(\mathbf{X}_s)\|_F^2 \right) \right]$$

$$= 2\eta^2 N\sigma^2 T \left[\frac{\tau_1 \zeta^{2\tau_2}}{1-\zeta^{2\tau_2}} + \tau_1 - 1 \right] +$$

$$2\eta^2 \beta\tau \sum_{r=0}^{\frac{T}{\tau}-1} \left[\left(\sum_{s=r\tau+1}^{r\tau+\tau_1} \|\nabla F(\mathbf{X}_s)\|_F^2 \right) \left(\sum_{j=r}^{\frac{T}{\tau}-1} \zeta^{2\tau_2(j-r)} \right) \right]$$

$$\leq 2\eta^2 N\sigma^2 T \left[\frac{\tau_1 \zeta^{2\tau_2}}{1-\zeta^{2\tau_2}} + \tau_1 - 1 \right] +$$

$$\frac{2\eta^2 \beta\tau}{1-\zeta^{2\tau_2}} \sum_{r=0}^{\frac{T}{\tau}-1} \sum_{s=r\tau+1}^{r\tau+\tau_1} \|\nabla F(\mathbf{X}_s)\|_F^2$$

$$= 2\eta^2 N\sigma^2 T \left(\frac{\tau_1}{1-\zeta^{2\tau_2}} - 1 \right) +$$

$$\frac{2\eta^2 \beta\tau}{1-\zeta^{2\tau_2}} \sum_{r=0}^{\frac{T}{\tau}-1} \sum_{s=r\tau+1}^{r\tau+\tau_1} \|\nabla F(\mathbf{X}_s)\|_F^2. \tag{84}$$

Thus we finish the first part for deriving the bound of A_1.

The bound of A_2: From the second term in (64), we have

$$A_2 = 2\eta^2 \mathbb{E} \left\| \sum_{r=0}^{j} \mathbf{Q}_r (\mathbf{C}^{(j-r)\tau_2} - \mathbf{J}) \right\|_F^2$$

$$= 2\eta^2 \sum_{r=0}^{j} \mathbb{E} \|\mathbf{Q}_r (\mathbf{C}^{(j-r)\tau_2} - \mathbf{J})\|_F^2 +$$

$$2\eta^2 \sum_{n=0}^{j} \sum_{l=0, l\neq n}^{j} \mathbb{E}[\mathrm{Tr}((\mathbf{C}^{\tau_2(j-n)} - \mathbf{J})\mathbf{Q}_n^\top \mathbf{Q}_l (\mathbf{C}^{\tau_2(j-l)} - \mathbf{J}))]. \tag{85}$$

Considering the second term in equality (85), we can find the bound of the trace as follows:

$$|\mathrm{Tr}((\mathbf{C}^{\tau_2(j-n)} - \mathbf{J})\mathbf{Q}_n^\top \mathbf{Q}_l (\mathbf{C}^{\tau_2(j-l)} - \mathbf{J}))|$$

$$\leq \|(\mathbf{C}^{\tau_2(j-n)} - \mathbf{J})\mathbf{Q}_n^\top\|_F \|\mathbf{Q}_l (\mathbf{C}^{\tau_2(j-l)} - \mathbf{J})\|_F \tag{86}$$

$$\leq \|\mathbf{C}^{\tau_2(j-n)} - \mathbf{J}\|_{op} \|\mathbf{Q}_n\|_F \|\mathbf{Q}_l\|_F \|\mathbf{C}^{\tau_2(j-l)} - \mathbf{J}\|_{op} \tag{87}$$

$$\leq \frac{1}{2} \zeta^{\tau_2(2j-n-l)} \left[\|\mathbf{Q}_n\|_F^2 + \|\mathbf{Q}_l\|_F^2 \right], \tag{88}$$

where (86) is because of Lemma 6, (87) comes from Lemma 5 and (88) is due to Lemma 7 and $2ab \leq a^2 + b^2$. Then we

have

$$A_2 \leq 2\eta^2 \sum_{r=0}^{j} \mathbb{E}\|\mathbf{Q}_r\|_F^2 \|\mathbf{C}^{\tau_2(j-r)} - \mathbf{J}\|_{op}^2 +$$

$$\eta^2 \sum_{n=0}^{j} \sum_{l=0, l\neq n}^{j} \zeta^{\tau_2(2j-n-l)} \mathbb{E}\left[\|\mathbf{Q}_n\|_F^2 + \|\mathbf{Q}_l\|_F^2 \right] \tag{89}$$

$$= 2\eta^2 \sum_{r=0}^{j} \zeta^{2\tau_2(j-r)} \mathbb{E}\|\mathbf{Q}_r\|_F^2 + 2\eta^2 \sum_{n=0}^{j} \sum_{l=0, l\neq n}^{j} \zeta^{\tau_2(2j-n-l)} \mathbb{E}\|\mathbf{Q}_n\|_F^2 \tag{90}$$

$$= 2\eta^2 \sum_{r=0}^{j} \zeta^{2\tau_2(j-r)} \mathbb{E}\|\mathbf{Q}_r\|_F^2 + 2\eta^2 \sum_{n=0}^{j} \zeta^{\tau_2(j-n)} \mathbb{E}\|\mathbf{Q}_n\|_F^2 \sum_{l=0, l\neq n}^{j} \zeta^{\tau_2(j-l)} \tag{91}$$

$$= 2\eta^2 \left[\sum_{r=0}^{j-1} \zeta^{2\tau_2(j-r)} \mathbb{E}\|\mathbf{Q}_r\|_F^2 + \sum_{n=0}^{j-1} \zeta^{\tau_2(j-n)} \mathbb{E}\|\mathbf{Q}_n\|_F^2 \sum_{l=0, l\neq n}^{j} \zeta^{\tau_2(j-l)} \right.$$

$$\left. + \mathbb{E}\|\mathbf{Q}_j\|_F^2 + \mathbb{E}\|\mathbf{Q}_j\|_F^2 \sum_{l=0}^{j-1} \zeta^{\tau_2(j-l)} \right] \tag{92}$$

$$\leq 2\eta^2 \left[\sum_{r=0}^{j-1} \zeta^{2\tau_2(j-r)} \mathbb{E}\|\mathbf{Q}_r\|_F^2 + \sum_{n=0}^{j-1} \frac{\zeta^{\tau_2(j-n)}}{1-\zeta^{\tau_2}} \mathbb{E}\|\mathbf{Q}_n\|_F^2 \right.$$

$$\left. + \mathbb{E}\|\mathbf{Q}_j\|_F^2 + \mathbb{E}\|\mathbf{Q}_j\|_F^2 \frac{\zeta^{\tau_2}}{1-\zeta^{\tau_2}} \right], \tag{93}$$

where (90) is because indices n and l are symmetric and (93) is due to

$$\sum_{l=0, l\neq n}^{j} \zeta^{\tau_2(j-l)} \leq \sum_{l=0}^{j} \zeta^{\tau_2 l} = \frac{1-\zeta^{\tau_2(j+1)}}{1-\zeta^{\tau_2}} < \frac{1}{1-\zeta^{\tau_2}}$$

$$\sum_{l=0}^{j-1} \zeta^{\tau_2(j-l)} = \sum_{l=1}^{j} \zeta^{\tau_2 l} = \frac{\zeta^{\tau_2}(1-\zeta^{\tau_2 j})}{1-\zeta^{\tau_2}} < \frac{\zeta^{\tau_2}}{1-\zeta^{\tau_2}}.$$

By rearranging (93), we obtain

$$A_2 \leq 2\eta^2 \sum_{r=0}^{j-1} \left[\left(\zeta^{2\tau_2(j-r)} + \frac{\zeta^{\tau_2(j-r)}}{1-\zeta^{\tau_2}} \right) \mathbb{E}\|\mathbf{Q}_r\|_F^2 \right] +$$

$$\frac{2\eta^2}{1-\zeta^{\tau_2}} \mathbb{E}\|\mathbf{Q}_j\|_F^2 \tag{94}$$

$$= 2\eta^2 \sum_{r=0}^{j-1} \left[\left(\zeta^{2\tau_2(j-r)} + \frac{\zeta^{\tau_2(j-r)}}{1-\zeta^{\tau_2}} \right) \mathbb{E}\left\| \sum_{s=1}^{\tau_1} \nabla F(\mathbf{X}_{r\tau+s}) \right\|_F^2 \right]$$

$$+ \frac{2\eta^2}{1-\zeta^{\tau_2}} \mathbb{E}\left\| \sum_{s=1}^{i-1} \nabla F(\mathbf{X}_{j\tau+s}) \right\|_F^2 \tag{95}$$

$$\leq 2\eta^2 \tau_1 \sum_{r=0}^{j-1} \left[\left(\zeta^{2\tau_2(j-r)} + \frac{\zeta^{\tau_2(j-r)}}{1-\zeta^{\tau_2}} \right) \sum_{s=1}^{\tau_1} \mathbb{E}\|\nabla F(\mathbf{X}_{r\tau+s})\|_F^2 \right]$$

$$+ \frac{2\eta^2(i-1)}{1-\zeta^{\tau_2}} \sum_{s=1}^{i-1} \mathbb{E}\|\nabla F(\mathbf{X}_{j\tau+s})\|_F^2, \tag{96}$$

where (96) is because of the convexity of Frobenius norm and Jensen's inequality. Then by summing over all iterations in

j-th round from $i = 1$ to $i = \tau$, we have

$$\sum_{i=1}^{\tau} A_2$$

$$\leq 2\eta^2 \tau_1 \tau \sum_{r=0}^{j-1} \left[\left(\zeta^{2\tau_2(j-r)} + \frac{\zeta^{\tau_2(j-r)}}{1 - \zeta^{\tau_2}} \right) \sum_{s=1}^{\tau_1} \mathbb{E}\|\nabla F(\mathbf{X}_{r\tau+s})\|_{\mathrm{F}}^2 \right]$$

$$+ \frac{\eta^2 \tau(\tau - 1)}{1 - \zeta^{\tau_2}} \sum_{s=1}^{\tau_1 - 1} \mathbb{E}\|\nabla F(\mathbf{X}_{j\tau+s})\|_{\mathrm{F}}^2. \tag{97}$$

Then, summing over all rounds from $j = 0$ to $j = \frac{T}{\tau} - 1$, where T is the total iterations, we can obtain

$$\sum_{j=0}^{\frac{T}{\tau}-1} \sum_{i=1}^{\tau} A_2$$

$$\leq 2\eta^2 \tau_1 \tau \sum_{j=0}^{\frac{T}{\tau}-1} \sum_{r=0}^{j-1} \left[\left(\zeta^{2\tau_2(j-r)} + \frac{\zeta^{\tau_2(j-r)}}{1 - \zeta^{\tau_2}} \right) \sum_{s=1}^{\tau_1} \mathbb{E}\|\nabla F(\mathbf{X}_{r\tau+s})\|_{\mathrm{F}}^2 \right]$$

$$+ \frac{\eta^2 \tau(\tau - 1)}{1 - \zeta^{\tau_2}} \sum_{j=0}^{\frac{T}{\tau}-1} \sum_{s=1}^{\tau_1 - 1} \mathbb{E}\|\nabla F(\mathbf{X}_{j\tau+s})\|_{\mathrm{F}}^2 \tag{98}$$

$$= 2\eta^2 \tau_1 \tau \sum_{r=0}^{\frac{T}{\tau}-2} \left[\left(\sum_{s=1}^{\tau_1} \mathbb{E}\|\nabla F(\mathbf{X}_{r\tau+s})\|_{\mathrm{F}}^2 \right) \sum_{j=1}^{\frac{T}{\tau}-1-r} \left(\zeta^{2\tau_2 j} + \frac{\zeta^{\tau_2 j}}{1 - \zeta^{\tau_2}} \right) \right]$$

$$+ \frac{\eta^2 \tau(\tau - 1)}{1 - \zeta^{\tau_2}} \sum_{j=0}^{\frac{T}{\tau}-1} \sum_{s=1}^{\tau_1 - 1} \mathbb{E}\|\nabla F(\mathbf{X}_{j\tau+s})\|_{\mathrm{F}}^2 \tag{99}$$

$$\leq 2\eta^2 \tau_1 \tau \left(\frac{\zeta^{2\tau_2}}{1 - \zeta^{2\tau_2}} + \frac{\zeta^{\tau_2}}{(1 - \zeta^{\tau_2})^2} \right) \sum_{r=0}^{\frac{T}{\tau}-2} \sum_{s=1}^{\tau_1} \mathbb{E}\|\nabla F(\mathbf{X}_{r\tau+s})\|_{\mathrm{F}}^2$$

$$+ \frac{\eta^2 \tau(\tau - 1)}{1 - \zeta^{\tau_2}} \sum_{j=0}^{\frac{T}{\tau}-1} \sum_{s=1}^{\tau_1 - 1} \mathbb{E}\|\nabla F(\mathbf{X}_{j\tau+s})\|_{\mathrm{F}}^2 \tag{100}$$

$$\leq \frac{\eta^2 \tau}{1 - \zeta^{\tau_2}} \left(\frac{2\tau_1 \zeta^{2\tau_2}}{1 + \zeta^{\tau_2}} + \frac{2\tau_1 \zeta^{\tau_2}}{1 - \zeta^{\tau_2}} + \tau - 1 \right) \sum_{r=0}^{\frac{T}{\tau}-1} \sum_{s=1}^{\tau_1} \mathbb{E}\|\nabla F(\mathbf{X}_{r\tau+s})\|_{\mathrm{F}}^2 \tag{101}$$

$$= \frac{\eta^2 \tau}{1 - \zeta^{\tau_2}} \left(\frac{2\tau_1 \zeta^{2\tau_2}}{1 + \zeta^{\tau_2}} + \frac{2\tau_1 \zeta^{\tau_2}}{1 - \zeta^{\tau_2}} + \tau - 1 \right) \sum_{r=0}^{\frac{T}{\tau}-1} \sum_{s=r\tau+1}^{r\tau+\tau_1} \mathbb{E}\|\nabla F(\mathbf{X}_s)\|_{\mathrm{F}}^2. \tag{102}$$

Therefore, we finish the second part for deriving the bound of A_2.

The final proof of Proposition 1: Based on inequalities

(64), (84) and (102), local drift has an upper bound as follows:

$$\frac{L^2}{TN} \sum_{t=1}^{T} \mathbb{E}\|\mathbf{X}_t(\mathbf{I} - \mathbf{J})\|_{\mathrm{F}}^2$$

$$\leq \frac{L^2}{TN} \sum_{j=0}^{\frac{T}{\tau}-1} \sum_{i=1}^{\tau} (A_1 + A_2) \tag{103}$$

$$\leq 2L^2 \eta^2 \sigma^2 \left(\frac{\tau_1}{1 - \zeta^{2\tau_2}} - 1 \right) + \frac{L^2}{TN} \frac{2\eta^2 \beta \tau}{1 - \zeta^{2\tau_2}} \sum_{r=0}^{\frac{T}{\tau}-1} \sum_{s=r\tau+1}^{r\tau+\tau_1} \|\nabla F(\mathbf{X}_s)\|_{\mathrm{F}}^2$$

$$+ \frac{L^2}{TN} \frac{\eta^2 \tau}{1 - \zeta^{\tau_2}} \left(\frac{2\tau_1 \zeta^{2\tau_2}}{1 + \zeta^{\tau_2}} + \frac{2\tau_1 \zeta^{\tau_2}}{1 - \zeta^{\tau_2}} + \tau - 1 \right) \sum_{r=0}^{\frac{T}{\tau}-1} \sum_{s=r\tau+1}^{r\tau+\tau_1} \mathbb{E}\|\nabla F(\mathbf{X}_s)\|_{\mathrm{F}}^2. \tag{104}$$

Substituting (104) into (51), we have

$$\mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \|\nabla F(\mathbf{u}_t)\|^2 \right] \leq \frac{2[F(\mathbf{u}_1) - F_{\inf}]}{\eta T} + \frac{\eta L \sigma^2}{N}$$

$$+ 2L^2 \eta^2 \sigma^2 \left(\frac{\tau_1}{1 - \zeta^{2\tau_2}} - 1 \right) + \frac{L^2}{TN} \frac{2\eta^2 \beta \tau}{1 - \zeta^{2\tau_2}} \sum_{r=0}^{\frac{T}{\tau}-1} \sum_{s=r\tau+1}^{r\tau+\tau_1} \|\nabla F(\mathbf{X}_s)\|_{\mathrm{F}}^2$$

$$+ \frac{L^2}{TN} \frac{\eta^2 \tau}{1 - \zeta^{\tau_2}} \left(\frac{2\tau_1 \zeta^{2\tau_2}}{1 + \zeta^{\tau_2}} + \frac{2\tau_1 \zeta^{\tau_2}}{1 - \zeta^{\tau_2}} + \tau - 1 \right) \sum_{r=0}^{\frac{T}{\tau}-1} \sum_{s=r\tau+1}^{r\tau+\tau_1} \mathbb{E}\|\nabla F(\mathbf{X}_s)\|_{\mathrm{F}}^2$$

$$- \frac{1}{NT} \left[1 - \eta L \left(\frac{\beta}{N} + 1 \right) \right] \sum_{t=1}^{T} \mathbb{E}\|\nabla F(\mathbf{X}_t)\|_{\mathrm{F}}^2. \tag{105}$$

Because there exists no gradient during inter-node communication period, i.e., $t \in [k]_2$, we have $\sum_{t=1}^{T} \mathbb{E}\|\nabla F(\mathbf{X}_t)\|_{\mathrm{F}}^2 = \sum_{r=0}^{\frac{T}{\tau}-1} \sum_{s=r\tau+1}^{r\tau+\tau_1} \mathbb{E}\|\nabla F(\mathbf{X}_s)\|_{\mathrm{F}}^2$. So if the learning rate satisfies the following inequality:

$$\eta L \left(\frac{\beta}{N} + 1 \right) + \frac{2\eta^2 L^2 \beta \tau}{1 - \zeta^{2\tau_2}} + \frac{\eta^2 L^2 \tau}{1 - \zeta^{\tau_2}} \left(\frac{2\tau_1 \zeta^{2\tau_2}}{1 + \zeta^{\tau_2}} + \frac{2\tau_1 \zeta^{\tau_2}}{1 - \zeta^{\tau_2}} + \tau - 1 \right) \leq 1, \tag{106}$$

we can obtain

$$\mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \|\nabla F(\mathbf{u}_t)\|^2 \right] \leq \frac{2[F(\mathbf{u}_1) - F_{\inf}]}{\eta T} + \frac{\eta L \sigma^2}{N}$$

$$+ 2L^2 \eta^2 \sigma^2 \left(\frac{\tau_1}{1 - \zeta^{2\tau_2}} - 1 \right). \tag{107}$$

Therefore, the proof of Proposition 1 is finished.

APPENDIX B
PROOF OF PROPOSITION 2

A. Preliminaries of Proof

Firstly, we present some basic definition and inequalities for the proof.

Remark 3. *We define* $\bar{\mathbf{X}}_t \triangleq \mathbf{X}_t \mathbf{J}$, *then* $\bar{\mathbf{X}}_t = [\mathbf{u}_t, ..., \mathbf{u}_t]$.

Remark 4. *If* $F(\cdot)$ *is L-smooth for parameter* $L \geq 0$, *we have*

$$F(\mathbf{y}) \leq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2,$$

$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Remark 5. *If $F(\cdot)$ is L-smooth with minimizer $\mathbf{x}^*$ (i.e., $\nabla F(\mathbf{x}^*) = 0$), then*

$$\|\nabla F(\mathbf{x})\|^2 = \|\nabla F(\mathbf{x}) - \nabla F(\mathbf{x}^*)\|^2 \leq 2L(F(\mathbf{x}) - F(\mathbf{x}^*)).$$

Remark 6. *If $F(\cdot)$ is μ-strong for parameter $\mu \geq 0$, we have*

$$F(\mathbf{y}) \geq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2,$$

$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Remark 7. *For $A \in \mathbb{R}^{d \times N}$, $B \in \mathbb{R}^{N \times N}$, $\|AB\|_F \leq \|A\|_F\|B\|_2$.*

Remark 8. *If $\mathbb{E}_{\xi|\mathbf{w}}\|g(\mathbf{w}^{(i)})\|^2 \leq G^2$, then for $\mathbf{G} = [g(\mathbf{w}^{(1)}), ..., g(\mathbf{w}^{(N)})]$, we have*

$$\mathbb{E}_{\xi_t^{(1)},...,\xi_t^{(N)}}\|\mathbf{G}\|_F^2 \leq NG^2.$$

Remark 9. *If for $F_i(\cdot)$ with $\mathbb{E}_{\xi|\mathbf{w}}\|g(\mathbf{w}^{(i)}) - \nabla F_i(\mathbf{w})\|^2 \leq \sigma_i^2$, then*

$$\mathbb{E}_{\xi_t^{(1)},...,\xi_t^{(N)}} \left\| \frac{1}{N}\sum_{i=1}^{N}\left(\nabla F_i(\mathbf{w}_t^{(i)}) - g(\mathbf{w}^{(i)})\right)\right\|^2 \leq \frac{\bar{\sigma}^2}{N},$$

where $\bar{\sigma}^2 = \frac{\sum_{i=1}^{N}\sigma_i^2}{N}$.

Proof. Assuming $Y_i = \nabla F_i(\mathbf{w}_t^{(i)}) - g(\mathbf{w}^{(i)})$, then

$$\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}Y_i\right\|^2 = \frac{1}{N^2}\left(\sum_{i=1}^{N}\mathbb{E}\|Y_i\|^2 + \sum_{i \neq j}\mathbb{E}\langle Y_i, Y_j \rangle\right)$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\|Y_i\|^2 \leq \frac{1}{N^2}\sum_{i=1}^{N}\sigma_i^2 = \frac{\bar{\sigma}^2}{N}.$$

Expectation of the product of Y_i and Y_j is zero because $\xi^{(i)}$ is independent of $\xi^{(j)}$ for $i \neq j$. $\square$

Remark 10. *For given two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$,*

$$\|\mathbf{a} + \mathbf{b}\|^2 \leq (1+\theta)\|\mathbf{a}\|^2 + (1+\theta^{-1})\|\mathbf{b}\|^2, \forall \theta > 0.$$

This inequality also holds for the sum of two matrices $A, B \in \mathbb{R}^{d \times N}$ in Frobenius norm.

B. The supporting lemmas

We present several supporting lemmas to prove Proposition 2 as follows. We define $\mathbf{u}^{(k)} \triangleq \mathbf{u}_{k\tau}$ for simplification, which is the average of model parameters after k iteration rounds.

Lemma 8. *If there exists $a_k \in \mathbb{R}$ which satisfies $\alpha_k(F(\mathbf{u}^{(k)}) - F^*) \geq \frac{1}{\tau_1}\sum_{j=0}^{\tau_1-1}(F(\mathbf{u}_{k\tau+j}) - F^*)$, then the averages $\mathbf{u}^{(k)}$ of Algorithm 2 satisfy*

$$\mathbb{E}_{\xi_{k\tau}^{(1)},...,\xi_{k\tau}^{(N)}}\|\mathbf{u}^{(k+1)} - \mathbf{u}^*\|^2 \leq \left(1 - \frac{\mu\eta_k\tau_1}{2}\right)\|\mathbf{u}^{(k)} - \mathbf{u}^*\|^2 +$$

$$\eta_k^2\tau_1^2\frac{\bar{\sigma}^2}{N} + \frac{(L+\mu)\eta_k^3\bar{\sigma}^2(\tau_1-1)(2\tau_1-1)\tau_1}{6N} +$$

$$(F(\mathbf{u}^{(k)}) - F^*)\left[4L\eta_k^2\tau_1^2\alpha_k - 2\eta_k\tau_1 + \frac{4}{3}(L+\mu)\eta_k^3L\alpha_k(\tau_1-1)(2\tau_1-1)\tau_1\right]$$

$$+ \frac{2L^2\eta_k^2\tau_1}{N}\Phi_k + \frac{2(L+\mu)\eta_k}{N}\Phi_k + \frac{(L+\mu)\eta_k}{N}\sum_{j=0}^{\tau_1-1}\sum_{i=1}^{N}\frac{4\eta_k^2L^2j^2}{N}\Phi_{k,j},$$

$$(108)$$

where $\Phi_k = \sum_{j=0}^{\tau_1-1}\sum_{i=1}^{N}\|\mathbf{u}_{k\tau+j} - \mathbf{w}_{k\tau+j}^{(i)}\|^2$ and $\Phi_{k,j} = \sum_{p=0}^{j-1}\sum_{i=1}^{N}\|\mathbf{u}_{k\tau+p} - \mathbf{w}_{k\tau+p}^{(i)}\|^2$.

Proof. According to the update of model averaging, we have

$$\|\mathbf{u}^{(k+1)} - \mathbf{u}^*\|^2$$

$$= \left\|\mathbf{u}^{(k)} - \mathbf{u}^* - \eta_k\sum_{j=0}^{\tau_1-1}\frac{1}{N}\sum_{q=1}^{N}g(\mathbf{w}_{k\tau+j}^{(q)})\right\|^2$$

$$= \left\|\mathbf{u}^{(k)} - \mathbf{u}^* - \sum_{j=0}^{\tau_1-1}\frac{\eta_k}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{w}_{k\tau+j}^{(i)}) + \right.$$

$$\left.\sum_{j=0}^{\tau_1-1}\frac{\eta_k}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{w}_{k\tau+j}^{(i)}) - \sum_{j=0}^{\tau_1-1}\frac{\eta_k}{N}\sum_{q=1}^{N}g(\mathbf{w}_{k\tau+j}^{(q)})\right\|^2$$

$$= \left\|\mathbf{u}^{(k)} - \mathbf{u}^* - \sum_{j=0}^{\tau_1-1}\frac{\eta_k}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{w}_{k\tau+j}^{(i)})\right\|^2 +$$

$$\left\|\sum_{j=0}^{\tau_1-1}\frac{\eta_k}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{w}_{k\tau+j}^{(i)}) - \sum_{j=0}^{\tau_1-1}\frac{\eta_k}{N}\sum_{q=1}^{N}g(\mathbf{w}_{k\tau+j}^{(q)})\right\|^2$$

$$+ \frac{2\eta_k}{N}\left\langle \mathbf{u}^{(k)} - \mathbf{u}^* - \sum_{j=0}^{\tau_1-1}\frac{\eta_k}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{w}_{k\tau+j}^{(i)}), \right.$$

$$\left.\sum_{j=0}^{\tau_1-1}\sum_{i=1}^{N}\nabla F_i(\mathbf{w}_{k\tau+j}^{(i)}) - \sum_{j=0}^{\tau_1-1}\sum_{q=1}^{N}g(\mathbf{w}_{k\tau+j}^{(q)})\right\rangle.$$

The last term is zero because $\mathbb{E}_{\xi^{(i)}|\mathbf{w}^{(i)}}[g(\mathbf{w}_t^{(i)})] = \nabla F(\mathbf{w}_t^{(i)})$. The second term is less than $\eta_k^2\tau_1^2\frac{\bar{\sigma}^2}{N}$. We rewrite the first term as

$$\left\|\mathbf{u}^{(k)} - \mathbf{u}^* - \sum_{j=0}^{\tau_1-1}\frac{\eta_k}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{w}_{k\tau+j}^{(i)})\right\|^2$$

$$= \|\mathbf{u}^{(k)} - \mathbf{u}^*\|^2 + \eta_k^2\underbrace{\left\|\sum_{j=0}^{\tau_1-1}\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{w}_{k\tau+j}^{(i)})\right\|^2}_{T_1}$$

$$- 2\eta_k\underbrace{\left\langle \mathbf{u}^{(k)} - \mathbf{u}^*, \sum_{j=0}^{\tau_1-1}\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\mathbf{w}_{k\tau+j}^{(i)})\right\rangle}_{T_2}.$$

Considering T_1, we have

$$
\begin{aligned}
T_1 &= \left\| \sum_{j=0}^{\tau_1-1} \frac{1}{N} \sum_{i=1}^{N} \left(\nabla F_i(\mathbf{w}_{k\tau+j}^{(i)}) - \nabla F_i(\mathbf{u}_{k\tau+j}) + \nabla F_i(\mathbf{u}_{k\tau+j}) \right. \right. \\
&\quad \left. \left. - \nabla F_i(\mathbf{w}^*) \right) \right\|^2 \\
&\leq 2\tau_1 \sum_{j=0}^{\tau_1} \left\| \frac{1}{N} \sum_{i=1}^{N} \left(\nabla F_i(\mathbf{w}_{k\tau+j}^{(i)}) - \nabla F_i(\mathbf{u}_{k\tau+j}) \right) \right\|^2 + \\
&\quad 2\tau_1 \sum_{j=0}^{\tau_1-1} \left\| \frac{1}{N} \sum_{i=1}^{N} (\nabla F_i(\mathbf{u}_{k\tau+j}) - \nabla F_i(\mathbf{w}^*)) \right\|^2 \\
&\leq \frac{2L^2}{N} \tau_1 \sum_{j=0}^{\tau_1-1} \sum_{i=1}^{N} \| \mathbf{w}_{k\tau+j}^{(i)} - \mathbf{u}_{k\tau+j} \|^2 \\
&\quad + 4L\tau_1 \sum_{j=0}^{\tau_1-1} (F(\mathbf{u}_{k\tau+j}) - F^*) \qquad (109) \\
&\leq \frac{2L^2}{N} \tau_1 \sum_{j=0}^{\tau_1-1} \sum_{i=1}^{N} \| \mathbf{w}_{k\tau+j}^{(i)} - \mathbf{u}_{k\tau+j} \|^2 + 4L\tau_1^2 \alpha_k (F(\mathbf{u}^{(k)}) - F^*). \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (110)
\end{aligned}
$$

According to condition 1 of Assumption 1 and Remark 5, we obtain (109). And (110) is due to the assumption of α_k: $\alpha_k(F(\mathbf{u}^{(k)}) - F^*) \geq \frac{1}{\tau_1} \sum_{j=0}^{\tau_1-1} (F(\mathbf{u}_{k\tau+j}) - F^*)$.

Then considering T_2, we have

$$
\begin{aligned}
-&\frac{1}{\eta_k} T_2 \\
=& -\frac{2}{N} \sum_{j=0}^{\tau_1-1} \sum_{i=1}^{N} \langle \mathbf{u}^{(k)} - \mathbf{u}^*, \nabla F_i(\mathbf{w}_{k\tau+j}^{(i)}) \rangle \\
=& -\frac{2}{N} \sum_{j=0}^{\tau_1-1} \sum_{i=1}^{N} \left[\langle \mathbf{u}^{(k)} - \mathbf{w}_{k\tau+j}^{(i)}, \nabla F_i(\mathbf{w}_{k\tau+j}^{(i)}) \rangle \right. \\
&\left. + \langle \mathbf{w}_{k\tau+j}^{(i)} - \mathbf{u}^*, \nabla F_i(\mathbf{w}_{k\tau+j}^{(i)}) \rangle \right] \\
\leq& -\frac{2}{N} \sum_{j=0}^{\tau_1-1} \sum_{i=1}^{N} \left[F_i(\mathbf{u}^{(k)}) - F_i(\mathbf{w}_{k\tau+j}^{(i)}) - \frac{L}{2} \| \mathbf{u}^{(k)} - \mathbf{w}_{k\tau+j}^{(i)} \|^2 \right. \\
&\left. + F_i(\mathbf{w}_{k\tau+j}^{(i)}) - F_i(\mathbf{u}^*) + \frac{\mu}{2} \| \mathbf{w}_{k\tau+j}^{(i)} - \mathbf{u}^* \|^2 \right] \qquad (111) \\
\leq& \sum_{j=0}^{\tau_1-1} \left[-2(F(\mathbf{u}^{(k)}) - F(\mathbf{u}^*)) + \frac{L+\mu}{N} \sum_{i=1}^{N} \| \mathbf{u}^{(k)} - \mathbf{w}_{k\tau+j}^{(i)} \|^2 \right. \\
&\left. - \frac{\mu}{2} \| \mathbf{u}^{(k)} - \mathbf{u}^* \|^2 \right] \qquad (112) \\
=& -2\tau_1 (F(\mathbf{u}^{(k)}) - F^*) + \frac{L+\mu}{N} \sum_{j=0}^{\tau_1-1} \sum_{i=1}^{N} \| \mathbf{u}^{(k)} - \mathbf{w}_{k\tau+j}^{(i)} \|^2 \\
&- \frac{\mu\tau_1}{2} \| \mathbf{u}^{(k)} - \mathbf{u}^* \|^2, \qquad (113)
\end{aligned}
$$

where (111) is because of Remark 4 and 6, and (112) is because $\|\sum_{i=1}^{n} \mathbf{a}_i\|^2 \leq n \sum_{i=1}^{n} \|\mathbf{a}_i\|^2$. Focusing on the second term of (113), we have

$$
\begin{aligned}
&\| \mathbf{u}^{(k)} - \mathbf{w}_{k\tau+j}^{(i)} \|^2 \\
=& \left\| \mathbf{u}_{k\tau+j} - \mathbf{w}_{k\tau+j}^{(i)} + \frac{\eta_k}{N} \sum_{p=0}^{j-1} \sum_{i=1}^{N} g(\mathbf{w}_{k\tau+p}^{(i)}) \right\|^2 \\
=& \left\| \mathbf{u}_{k\tau+j} - \mathbf{w}_{k\tau+j}^{(i)} + \frac{\eta_k}{N} \sum_{p=0}^{j-1} \sum_{i=1}^{N} g(\mathbf{w}_{k\tau+p}^{(i)}) \right. \\
&\left. - \frac{\eta_k}{N} \sum_{p=0}^{j-1} \sum_{i=1}^{N} \nabla F_i(\mathbf{w}_{k\tau+p}^{(i)}) + \frac{\eta_k}{N} \sum_{p=0}^{j-1} \sum_{i=1}^{N} \nabla F_i(\mathbf{w}_{k\tau+p}^{(i)}) \right\|^2 \\
=& \left\| \mathbf{u}_{k\tau+j} - \mathbf{w}_{k\tau+j}^{(i)} + \frac{\eta_k}{N} \sum_{p=0}^{j-1} \sum_{i=1}^{N} \nabla F_i(\mathbf{w}_{k\tau+p}^{(i)}) \right\|^2 + \\
&\eta_k^2 \left\| \frac{1}{N} \sum_{p=0}^{j-1} \sum_{i=1}^{N} (g(\mathbf{w}_{k\tau+p}^{(i)}) - \nabla F_i(\mathbf{w}_{k\tau+p}^{(i)})) \right\|^2 \\
\leq& \left\| \mathbf{u}_{k\tau+j} - \mathbf{w}_{k\tau+j}^{(i)} + \frac{\eta_k}{N} \sum_{p=0}^{j-1} \sum_{i=1}^{N} \nabla F_i(\mathbf{w}_{k\tau+p}^{(i)}) \right\|^2 + \frac{\eta_k^2 j^2 \bar{\sigma}^2}{N} \\
\leq& 2\| \mathbf{u}_{k\tau+j} - \mathbf{w}_{k\tau+j}^{(i)} \|^2 + 2\eta_k^2 \left\| \frac{1}{N} \sum_{p=0}^{j-1} \sum_{i=1}^{N} \nabla F_i(\mathbf{w}_{k\tau+p}^{(i)}) \right\|^2 \\
&+ \frac{\eta_k^2 j^2 \bar{\sigma}^2}{N}. \qquad (114)
\end{aligned}
$$

Because the second term in the last inequality has a similar form with T_1, then following (110), we can obtain

$$
\begin{aligned}
&\left\| \frac{1}{N} \sum_{p=0}^{j-1} \sum_{i=1}^{N} \nabla F_i(\mathbf{w}_{k\tau+p}^{(i)}) \right\|^2 \\
\leq& \frac{2L^2}{N} j \sum_{p=0}^{j-1} \sum_{i=1}^{N} \| \mathbf{w}_{k\tau+p}^{(i)} - \mathbf{u}_{k\tau+p} \|^2 + 4Lj^2 \alpha_k (F(\mathbf{u}^{(k)}) - F^*). \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (115)
\end{aligned}
$$

By substituting (115) into (114), we have

$$
\begin{aligned}
&\| \mathbf{u}^{(k)} - \mathbf{w}_{k\tau+j}^{(i)} \|^2 \\
\leq& 2\| \mathbf{u}_{k\tau+j} - \mathbf{w}_{k\tau+j}^{(i)} \|^2 + \frac{4\eta_k^2 L^2 j}{N} \sum_{p=0}^{j-1} \sum_{i=1}^{N} \| \mathbf{w}_{k\tau+p}^{(i)} - \mathbf{u}_{k\tau+p} \|^2 + \\
&8\eta_k^2 L j^2 \alpha_k (F(\mathbf{u}^{(k)}) - F^*) + \frac{\eta_k^2 j^2 \bar{\sigma}^2}{N}. \qquad (116)
\end{aligned}
$$

Then by putting (116) into (113), we have the upper bound of

$-\frac{1}{\eta_k}T_2$ as

$$-\frac{1}{\eta_k}T_2 \leq$$

$$-2\tau_1(F(\mathbf{u}^{(k)}) - F^*) + \frac{L+\mu}{N}\sum_{j=0}^{\tau_1-1}\sum_{i=1}^{N}\|\mathbf{u}^{(k)} - \mathbf{w}_{k\tau+j}^{(i)}\|^2$$

$$-\frac{\mu\tau_1}{2}\|\mathbf{u}^{(k)} - \mathbf{u}^*\|^2$$

$$\leq -2\tau_1(F(\mathbf{u}^{(k)}) - F^*) + \frac{2(L+\mu)}{N}\sum_{j=0}^{\tau_1-1}\sum_{i=1}^{N}\|\mathbf{u}_{k\tau+j} - \mathbf{w}_{k\tau+j}^{(i)}\|^2$$

$$+\frac{L+\mu}{N}\sum_{j=0}^{\tau_1-1}\sum_{i=1}^{N}\frac{4\eta_k^2 L^2 j}{N}\sum_{p=0}^{j-1}\sum_{i=1}^{N}\|\mathbf{w}_{k\tau+p}^{(i)} - \mathbf{u}_{k\tau+p}\|^2+$$

$$\frac{4}{3}(L+\mu)L\eta_k^2\alpha_k(\tau_1-1)(2\tau_1-1)\tau_1(F(\mathbf{u}^{(k)}) - F^*)+$$

$$\frac{(L+\mu)\eta_k^2\bar{\sigma}^2(\tau_1-1)(2\tau_1-1)\tau_1}{6N} - \frac{\mu\tau_1}{2}\|\mathbf{u}^{(k)} - \mathbf{u}^*\|^2. \tag{117}$$

Therefore, we can finish the proof by putting the upper bounds of T_1 and T_2 into the original inequality as follows.

$$\|\mathbf{u}^{(k+1)} - \mathbf{u}^*\|^2$$

$$\leq \|\mathbf{u}^{(k)} - \mathbf{u}^*\|^2 + \frac{2L^2\eta_k^2}{N}\tau_1\sum_{j=0}^{\tau_1-1}\sum_{i=1}^{N}\|\mathbf{w}_{k\tau+j}^{(i)} - \mathbf{u}_{k\tau+j}\|^2$$

$$+ 4\eta_k^2 L\tau_1^2\alpha_k(F(\mathbf{u}^{(k)}) - F^*) \boxed{-2\eta_k}\tau_1(F(\mathbf{u}^{(k)}) - F^*)+$$

$$\frac{2(L+\mu)\eta_k}{N}\sum_{j=0}^{\tau_1-1}\sum_{i=1}^{N}\|\mathbf{u}_{k\tau+j} - \mathbf{w}_{k\tau+j}^{(i)}\|^2+$$

$$\frac{(L+\mu)\eta_k}{N}\sum_{j=0}^{\tau_1-1}\sum_{i=1}^{N}\frac{4\eta_k^2 L^2 j}{N}\sum_{p=0}^{j-1}\sum_{i=1}^{N}\|\mathbf{w}_{k\tau+p}^{(i)} - \mathbf{u}_{k\tau+p}\|^2+$$

$$\frac{4}{3}(L+\mu)L\eta_k^3\alpha_k(\tau_1-1)(2\tau_1-1)\tau_1(F(\mathbf{u}^{(k)}) - F^*)+$$

$$\frac{(L+\mu)\eta_k^3\bar{\sigma}^2(\tau_1-1)(2\tau_1-1)\tau_1}{6N} - \frac{\mu\eta_k\tau_1}{2}\|\mathbf{u}^{(k)} - \mathbf{u}^*\|^2$$

$$+ \eta_k^2\tau_1^2\frac{\bar{\sigma}^2}{N}$$

$$= (1 - \frac{\mu\eta_k\tau_1}{2})\|\mathbf{u}^{(k)} - \mathbf{u}^*\|^2 + \eta_k^2\tau_1^2\frac{\bar{\sigma}^2}{N}+$$

$$\frac{(L+\mu)\eta_k^3\bar{\sigma}^2(\tau_1-1)(2\tau_1-1)\tau_1}{6N}+$$

$$(F(\mathbf{u}^{(k)}) - F^*)\left[-2\eta_k\tau_1 + 4\eta_k^2 L\tau_1^2\alpha_k+\right.$$

$$\left.\frac{4}{3}(L+\mu)L\eta_k^3\alpha_k(\tau_1-1)(2\tau_1-1)\tau_1\right]+$$

$$\frac{2(L+\mu)\eta_k}{N}\sum_{j=0}^{\tau_1-1}\sum_{i=1}^{N}\|\mathbf{u}_{k\tau+j} - \mathbf{w}_{k\tau+j}^{(i)}\|^2+$$

$$\frac{2L^2\eta_k^2}{N}\tau_1\sum_{j=0}^{\tau_1-1}\sum_{i=1}^{N}\|\mathbf{w}_{k\tau+j}^{(i)} - \mathbf{u}_{k\tau+j}\|^2+$$

$$\frac{(L+\mu)\eta_k}{N}\sum_{j=0}^{\tau_1-1}\sum_{i=1}^{N}\frac{4\eta_k^2 L^2 j}{N}\sum_{p=0}^{j-1}\sum_{i=1}^{N}\|\mathbf{w}_{k\tau+p}^{(i)} - \mathbf{u}_{k\tau+p}\|^2.$$

$\square$

We define the function $h : \mathbb{R}^{d\times N} \times \mathbb{R}^{d\times N} \rightarrow \mathbb{R}^{d\times N} \times \mathbb{R}^{d\times N}$ denotes the average consensus operator, where $(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1}) = h(\mathbf{X}_t, \mathbf{Y}_t)$. Therefore, for CHOCO-G algorithm which is used in C-DFL, the operator h is

$$\mathbf{X}_{t+1} = \mathbf{X}_t + \gamma\mathbf{Y}_t(\mathbf{C} - \mathbf{I}),$$
$$\mathbf{Y}_{t+1} = \mathbf{Y}_t + Q(\mathbf{X}_{t+1} - \mathbf{Y}_t),$$

where $\mathbf{Y}_t = \hat{\mathbf{X}}_t$. For convenience, we use $\mathbf{X}^{(k)}, \bar{\mathbf{X}}^{(k)}$ to denote $\mathbf{X}_{k\tau}, \bar{\mathbf{X}}_{k\tau}$, respectively.

Lemma 9 ([34]). *CHOCO-G converges linearly for average consensus:*

$$e_t = (1-p)^t e_0,$$

where $p = \frac{\rho^2\delta}{82}$, where δ is the compression ratio as in Assumption 2, $e_t = \mathbb{E}_Q\sum_{i=1}^{N}\left(\|\mathbf{w}_t^{(i)} - \mathbf{u}_t\|^2 + \|\mathbf{w}_t^{(i)} - \hat{\mathbf{w}}_t^{(i)}\|^2\right)$ and using the step size $\gamma = \frac{\rho^2\delta}{16\rho+\rho^2+4\beta^2+2\rho\beta^2-8\rho\delta}$.

From Lemma 9, we have

$$\mathbb{E}\left[\|\mathbf{X}_{t+1} - \bar{\mathbf{X}}_{t+1}\|_F^2 + \|\mathbf{X}_{t+1} - \mathbf{Y}_{t+1}\|_F^2\right] \leq$$
$$(1-p)\mathbb{E}\left[\|\mathbf{X}_t - \bar{\mathbf{X}}_t\|_F^2 + \|\mathbf{X}_t - \mathbf{Y}_t\|_F^2\right]. \tag{118}$$

Lemma 10. *The iterates of C-DFL $\{\mathbf{X}^{(k)}\}$ satisfy*

$$\left\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_F^2$$
$$\leq \frac{2(1+\theta^{-1})\eta_k^2\tau_1^2 NG^2(1-p)^{\tau_2}}{(1+\theta)(1-p)^{\tau_2} - 1}[((1+\theta)(1-p)^{\tau_2})^k - 1].$$

Proof. Using (118) which shows the linear convergence of CHOCO-G, we have

$$\mathbb{E}\left[\left\|\mathbf{X}^{(k+1)} - \bar{\mathbf{X}}^{(k+1)}\right\|_F^2 + \left\|\mathbf{X}^{(k+1)} - \mathbf{Y}^{(k+1)}\right\|_F^2\right] \leq$$

$$(1-p)^{\tau_2}\mathbb{E}\left\|\mathbf{X}_{k\tau+\tau_1} - \bar{\mathbf{X}}_{k\tau+\tau_1}\right\|_F^2 + (1-p)^{\tau_2}\mathbb{E}\left\|\mathbf{X}_{k\tau+\tau_1} - \mathbf{Y}_{k\tau}\right\|_F^2$$

$$= (1-p)^{\tau_2}\mathbb{E}\left\|\mathbf{X}_{k\tau} - \bar{\mathbf{X}}_{k\tau} + \eta_k\sum_{j=0}^{\tau_1-1}\mathbf{G}_{k\tau+j}(\mathbf{J} - \mathbf{I})\right\|_F^2 +$$

$$(1-p)^{\tau_2}\mathbb{E}\left\|\mathbf{X}_{k\tau} - \mathbf{Y}_{k\tau} - \eta_k\sum_{j=0}^{\tau_1-1}\mathbf{G}_{k\tau+j}\right\|_F^2$$

$$\leq (1+\theta)(1-p)^{\tau_2}\mathbb{E}\left(\|\mathbf{X}_{k\tau} - \bar{\mathbf{X}}_{k\tau}\|_F^2 + \|\mathbf{X}_{k\tau} - \mathbf{Y}_{k\tau}\|_F^2\right)+$$

$$(1+\theta^{-1})(1-p)^{\tau_2}\eta_k^2\mathbb{E}\left(\left\|\sum_{j=0}^{\tau_1-1}\mathbf{G}_{k\tau+j}(\mathbf{J} - \mathbf{I})\right\|_F^2 + \left\|\sum_{j=0}^{\tau_1-1}\mathbf{G}_{k\tau+j}\right\|_F^2\right) \tag{119}$$

$$\leq (1+\theta)(1-p)^{\tau_2}\mathbb{E}\left(\|\mathbf{X}_{k\tau} - \bar{\mathbf{X}}_{k\tau}\|_F^2 + \|\mathbf{X}_{k\tau} - \mathbf{Y}_{k\tau}\|_F^2\right)$$
$$+ 2(1+\theta^{-1})(1-p)^{\tau_2}\eta_k^2\tau_1^2 NG^2, \tag{120}$$

where (119) follows from Remark 10 and (120) is from Remark 7 and 8. For sequence $\{e_k\}$, if $e_{k+1} \leq Ae_k + B$ where A and B are constants, then we can easily obtain

$$e_k \leq \frac{B}{A-1}(A^k - 1),$$

when $e_0 = 0$. We set $e_k = \mathbb{E}\left(\left\|\mathbf{X}_{k\tau} - \bar{\mathbf{X}}_{k\tau}\right\|_F^2 + \left\|\mathbf{X}_{k\tau} - \mathbf{Y}_{k\tau}\right\|_F^2\right)$ with $e_0 = 0$, $A = 2(1-p)^{\tau_2}$ and $B = 4(1-p)^{\tau_2}\eta_k^2\tau_1^2NG^2$. Then

$$\mathbb{E}\left(\left\|\mathbf{X}_{k\tau} - \bar{\mathbf{X}}_{k\tau}\right\|_F^2 + \left\|\mathbf{X}_{k\tau} - \mathbf{Y}_{k\tau}\right\|_F^2\right)$$
$$\leq \frac{2(1+\theta^{-1})\eta_k^2\tau_1^2NG^2(1-p)^{\tau_2}}{(1+\theta)(1-p)^{\tau_2}-1}[((1+\theta)(1-p)^{\tau_2})^k-1].$$

We finish the proof of Lemma 10 from

$$\mathbb{E}\left(\left\|\mathbf{X}^{(k)} - \bar{\mathbf{X}}^{(k)}\right\|_F^2\right) \leq \mathbb{E}\left(\left\|\mathbf{X}_{k\tau} - \bar{\mathbf{X}}_{k\tau}\right\|_F^2 + \left\|\mathbf{X}_{k\tau} - \mathbf{Y}_{k\tau}\right\|_F^2\right).$$

$\square$

Lemma 11. *Let* $\Phi_k = \sum_{j=0}^{\tau_1-1}\sum_{i=1}^{N}\|\mathbf{u}_{k\tau+j} - \mathbf{w}_{k\tau+j}^{(i)}\|^2$ *and* $\Phi_{k,j} = \sum_{p=0}^{j-1}\sum_{i=1}^{N}\|\mathbf{u}_{k\tau+p} - \mathbf{w}_{k\tau+p}^{(i)}\|^2$. *We have*

$$\Phi_k \leq \frac{4(1+\theta^{-1})\eta_k^2\tau_1^3NG^2(1-p)^{\tau_2}}{(1+\theta)(1-p)^{\tau_2}-1}[((1+\theta)(1-p)^{\tau_2})^k-1]$$
$$+ \frac{1}{3}\eta_k^2NG^2(\tau_1-1)(2\tau_1-1)\tau_1, \tag{121}$$

$$\Phi_{k,j} \leq \frac{4(1+\theta^{-1})\eta_k^2j^3NG^2(1-p)^{\tau_2}}{(1+\theta)(1-p)^{\tau_2}-1}[((1+\theta)(1-p)^{\tau_2})^k-1]$$
$$+ \frac{1}{3}\eta_k^2NG^2(j-1)(2j-1)j. \tag{122}$$

Proof. Straightforwardly, We have

$$\left\|\mathbf{X}_{k\tau+j} - \bar{\mathbf{X}}_{k\tau+j}\right\|_F^2$$
$$= \left\|\mathbf{X}_{k\tau} - \bar{\mathbf{X}}_{k\tau} + \eta_k\sum_{p=0}^{j-1}\mathbf{G}_{k\tau+p}(\mathbf{J}-\mathbf{I})\right\|_F^2$$
$$\leq 2\left\|\mathbf{X}_{k\tau} - \bar{\mathbf{X}}_{k\tau}\right\|_F^2 + 2\eta_k^2\left\|\sum_{p=0}^{j-1}\mathbf{G}_{k\tau+p}(\mathbf{J}-\mathbf{I})\right\|_F^2$$
$$\leq 2\left\|\mathbf{X}_{k\tau} - \bar{\mathbf{X}}_{k\tau}\right\|_F^2 + 2\eta_k^2j^2NG^2.$$

Therefore, we have

$$\Phi_k = \sum_{j=0}^{\tau_1}\left\|\mathbf{X}_{k\tau+j} - \bar{\mathbf{X}}_{k\tau+j}\right\|_F^2$$
$$\leq 2\sum_{j=0}^{\tau_1}\left\|\mathbf{X}_{k\tau} - \bar{\mathbf{X}}_{k\tau}\right\|_F^2 + 2\eta_k^2NG^2\sum_{j=0}^{\tau_1}j^2$$
$$\leq \frac{4(1+\theta^{-1})\eta_k^2\tau_1^3NG^2(1-p)^{\tau_2}}{(1+\theta)(1-p)^{\tau_2}-1}[((1+\theta)(1-p)^{\tau_2})^k-1]$$
$$+ \frac{1}{3}\eta_k^2NG^2(\tau_1-1)(2\tau_1-1)\tau_1,$$

where the last inequality follows from Lemma 10. Similarly, we can prove (122) by following the way of proving (121). $\square$

Lemma 12 ([35]). *Let* $\{a_t\}_{t\geq 0}$, $a_t \leq 0$, $\{e_t\}_{t\geq 0}$, $e_t \geq 0$, *be sequences satisfying*

$$a_{t+1} \leq (1-\mu\eta_t)a_t - \eta_te_tA + \eta_t^2B + \eta_t^3C,$$

for $\eta_t = \frac{4}{\mu(a+t)}$ *and constants* $A > 0$, B, $C \geq 0$, $\mu > 0$, $a > 1$. *Then*

$$\frac{A}{S_T}\sum_{t=0}^{T-1}w_te_t \leq \frac{\mu a^3}{4S_T}a_0 + \frac{2T(T+2a)}{\mu S_T}B + \frac{16T}{\mu^2S_T}C,$$

for $w_t = (a+t)^2$ *and* $S_T := \sum_{t=0}^{T-1}w_t = \frac{T}{6}(2T^2 + 6aT - 3T + 6a^2 - 6a + 1) \geq \frac{1}{3}T^3$.

C. Proof of Proposition 2

Lemma 13 (Convergence of C-DFL). *Under Assumption 1 and 2 for* $p = \frac{\rho^2\delta}{82}$, *if*

$$4L\eta_k^2\alpha_k - 2\eta_k\tau_1 + \frac{8}{3}(L+\mu)\eta_k^3L\alpha_k\tau_1^3 \leq -\eta_k,$$

Algorithm 2 with step size $\eta_k = \frac{4}{\mu(a+k)}$, *for parameter* $a \geq 16\kappa$, $\kappa = \frac{L}{\mu}$ *converges at rate*

$$F(\mathbf{w}_{avg}^{(K)}) - F^* \leq \frac{\mu\tau_1a^3}{8S_K}\|\mathbf{u}_0 - \mathbf{u}^*\|^2 + \frac{4K(K+2a)}{\mu S_K}\frac{\tau_1\bar{\sigma}^2}{N} + \frac{64K}{\mu^2S_K}\left(\frac{(L+\mu)\bar{\sigma}^2\tau_1}{3N} + D_1 + D_2 + D_3\right),$$

where $\mathbf{w}_{avg}^{(K)} = \frac{1}{S_K}\sum_{k=0}^{K}w_k\mathbf{u}_k$ *for weights* $w_k = (a+k)^2$, $S_K = \sum_{k=0}^{K}w_k \geq \frac{1}{3}T^2$, *and*

$$D_1 = \frac{2(1+\theta^{-1})L\tau_1^2G^2(1-p)^{\tau_2}}{(1+\theta)(1-p)^{\tau_2}-1}[((1+\theta)(1-p)^{\tau_2})^k-1]$$
$$+ \frac{1}{3}L\tau_1^2G^2,$$
$$D_2 = \frac{8(1+\theta^{-1})(L+\mu)\tau_1G^2(1-p)^{\tau_2}}{(1+\theta)(1-p)^{\tau_2}-1}[((1+\theta)(1-p)^{\tau_2})^k-1]$$
$$+ \frac{4}{3}(L+\mu)\tau_1G^2,$$
$$D_3 = \frac{(1+\theta^{-1})(L+\mu)\tau_1^4G^2(1-p)^{\tau_2}}{6((1+\theta)(1-p)^{\tau_2}-1)}[((1+\theta)(1-p)^{\tau_2})^k-1]$$
$$+ \frac{1}{36}(L+\mu)\tau_1^4G^2.$$

Proof. Combining the results of lemma 8 and 11, by using a_k

to denote $\mathbb{E}\|\mathbf{u}^{(k)} - \mathbf{u}^*\|^2$, i.e., $a_k := \mathbb{E}\|\mathbf{u}^{(k)} - \mathbf{u}^*\|^2$, we have

$$a_{k+1} \leq \left(1 - \frac{\mu\eta_k\tau_1}{2}\right)a_k + \frac{\eta_k^2\tau_1^2\bar{\sigma}^2}{N} +$$

$$(F(\mathbf{u}^{(k)}) - F^*)\left[-2\eta_k\tau_1 + 4\eta_k^2 L\tau_1^2\alpha_k + \frac{8}{3}(L+\mu)L\eta_k^3\alpha_k\tau_1^3\right]$$

$$+\eta_k^3\left[\frac{(L+\mu)\bar{\sigma}^2\tau_1^3}{3N}\right.$$

$$+\frac{8(1+\theta^{-1})L^2\eta_k\tau_1^4 G^2(1-p)^{\tau_2}}{(1+\theta)(1-p)^{\tau_2}-1}(((1+\theta)(1-p)^{\tau_2})^k-1)$$

$$+\frac{4}{3}L^2\eta_k\tau_1^4 G^2 +$$

$$\frac{8(1+\theta^{-1})(L+\mu)\tau_1^3 G^2(1-p)^{\tau_2}}{(1+\theta)(1-p)^{\tau_2}-1}(((1+\theta)(1-p)^{\tau_2})^k-1)$$

$$+\frac{4}{3}(L+\mu)\tau_1^3 G^2 +$$

$$\frac{8(1+\theta^{-1})L^2(L+\mu)\eta_k^2\tau_1^6 G^2(1-p)^{\tau_2}}{3(1+\theta)(1-p)^{\tau_2}-1}(((1+\theta)(1-p)^{\tau_2})^k-1)$$

$$\left.+\frac{4}{9}L^2(L+\mu)\eta_k^2\tau_1^6 G^2\right] \tag{123}$$

$$\leq \left(1 - \frac{\mu\eta_k\tau_1}{2}\right)a_k + \frac{\eta_k^2\tau_1^2\bar{\sigma}^2}{N} +$$

$$(F(\mathbf{u}^{(k)}) - F^*)\left[-2\eta_k\tau_1 + 4\eta_k^2 L\tau_1^2\alpha_k + \frac{8}{3}(L+\mu)L\eta_k^3\alpha_k\tau_1^3\right]$$

$$+ \eta_k^3\left[\frac{(L+\mu)\bar{\sigma}^2\tau_1^3}{3N} + \right.$$

$$\frac{2(1+\theta^{-1})L\tau_1^4 G^2(1-p)^{\tau_2}}{(1+\theta)(1-p)^{\tau_2}-1}(((1+\theta)(1-p)^{\tau_2})^k - 1)$$

$$+\frac{1}{3}L\tau_1^4 G^2 +$$

$$\frac{8(1+\theta^{-1})(L+\mu)\tau_1^3 G^2(1-p)^{\tau_2}}{(1+\theta)(1-p)^{\tau_2}-1}(((1+\theta)(1-p)^{\tau_2})^k-1)$$

$$+\frac{4}{3}(L+\mu)\tau_1^3 G^2 +$$

$$\frac{(1+\theta^{-1})(L+\mu)\tau_1^6 G^2(1-p)^{\tau_2}}{6((1+\theta)(1-p)^{\tau_2}-1)}(((1+\theta)(1-p)^{\tau_2})^k-1)$$

$$\left.+\frac{1}{36}(L+\mu)\tau_1^6 G^2\right], \tag{124}$$

where inequality (123) holds because $(\tau_1-1)(2\tau_1-1)\tau_1 \leq 2\tau_1^3$,

$$\sum_{j=0}^{\tau_1-1} j^5 = \frac{(\tau_1-1)^2\tau_1^2(2\tau^2-2\tau_1-1)}{12} \leq \frac{\tau_1^6}{6},$$

which is in the last two terms, and (124) holds because $\eta_k L \leq \frac{1}{4}$ (this holds, as $a \geq 16\kappa$). From Lemma 12, we prove this lemma. $\square$

Proof of Proposition 2. The proof of Proposition 2 follows from Lemma 13 with using the inequality $\mathbb{E}\mu\|\mathbf{u}_0 - \mathbf{u}^*\| \leq 2G$ derived in [40, Lemma 2] for the upper bound of the first term.

APPENDIX C
STRUCTURE DESCRIPTION OF CNNS

In this part, we present the structure of the two CNN models which are trained based on MNIST and CIFAR-10,

respectively. Table II shows the CNN models applied in the simulation.

TABLE II: The structure of CNNs

CNN trained on MNIST	CNN trained on CIFAR-10
Convolution layer[1][1,16,3×3]	Convolution layer [3,64,5×5]
ReLU layer	ReLU layer
MaxPool layer[2][2×2]	MaxPool layer [3×3]
Convolution layer [16,32,3×3]	Convolution layer [64,64,5×5]
ReLU layer	ReLU layer
MaxPool layer [2×2]	MaxPool layer [3×3]
Dense layer[3][32×7×7,10]	Dense layer [64×4×4,384]
	ReLU layer
	Dense layer [384,192]
	ReLU layer
	Dense layer [192,10]

[1] Convolution layer [input channels, output channels, kernel size]
[2] Maxpool layer [kernel size]
[3] Dense layer [size of each input sample, size of each output sample]