This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2023.3296624

SAFARI: Sparsity-Enabled Federated Learning with Limited and Unreliable Communications

Yuzhu Mao*, Zihao Zhao*, Meilin Yang, Le Liang, Yang Liu, Wenbo Ding,
Tian Lan, *Senior Member, IEEE,* and Xiao-Ping Zhang, *Fellow, IEEE*

Abstract—Federated learning (FL) enables edge devices to collaboratively learn a model in a distributed fashion. Many existing researches have focused on improving communication efficiency of high-dimensional models and addressing bias caused by local updates. However, most FL algorithms are either based on reliable communications or assuming fixed and known unreliability characteristics. In practice, networks could suffer from dynamic channel conditions and non-deterministic disruptions, with time-varying and unknown characteristics. To this end, in this paper we propose a sparsity-enabled FL framework with both improved communication efficiency and bias reduction, termed as SAFARI. It makes use of similarity among client models to rectify and compensate for bias that results from unreliable communications. More precisely, sparse learning is implemented on local clients to mitigate communication overhead, while to cope with unreliable communications, a similarity-based compensation method is proposed to provide surrogates for missing model updates. With respect to sparse models, we analyze SAFARI under bounded dissimilarity. It is demonstrated that SAFARI under unreliable communications is guaranteed to converge at the same rate as the standard FedAvg with perfect communications. Implementations and evaluations on the CIFAR-10 dataset validate the effectiveness of SAFARI by showing that it can achieve the same convergence speed and accuracy as FedAvg with perfect communications, with up to 60% of the model weights being pruned and a high percentage of client updates missing in each round of model updates.

Index Terms—Distributed networks, federated learning

1 Introduction

With rapid deployment of mobile sensing and computing devices, there are growing interests in fully exploiting distributed computing resources, as well as huge volumes of data generated at the network edge, for efficient learning [1]. To this end, federated learning (FL) [2], [3] enables distributed edge devices to collaboratively learn a model while maintaining data privacy [4]–[6], by allowing a central server and distributed clients to exchange updated model parameters and perform global aggregations. As wireless communications in practice often have limited network capacity [1], [2], a number of proposals have been made on communication-efficient FL. Examples include model pruning and sparsity-enabled design to exploit the structural redundancy of dense models [7] and performing local training for multiple epochs before periodic global aggregation in order to mitigate communication overhead [8]–[10].

Nevertheless, most existing FL algorithms either are based on reliable communications [9], [10] or assume fixed and known unreliability characteristics [11], [12]. These assumptions may not hold in real-world FL applications. Protocols for data-intensive communications like the lightweight User Datagram Protocol (UDP) tend to focus on best effort delivery without mechanisms for detecting failures and re-transmission. Reliable transmission of local updates cannot be guaranteed [11]. Further, an underlying wireless network could suffer from dynamic channel conditions and non-deterministic disruptions, whose characteristics are often unknown and time-varying. This raises serious challenges in FL – unpredictable absence of local updates with time-varying characteristics would lead to non-homogeneous bias under non independent and identically distributed (non-IID) data distribution, potentially introducing an unknown drift and causing slow and unstable convergence.

In this paper, we propose a Sparsity-enAbled Federated leArning framework under limited and unReliable communIcations, termed as SAFARI. When unreliability characteristics are unknown and potentially time-varying, we show that it is possible to rectify the resulting bias in global model aggregation by leveraging similarity among different client models. More precisely, once distributed clients locally train their models with sparse algorithms, the central server (i) updates a similarity matrix tracking the similarity among different clients based on received sparse models, and (ii) for any absent update in the current round, substitutes it with an available update received from the most similar client. Intuitively, these similarity-based surrogates provide an optimal way of compensating for any missing local

* *These authors contribute equally.*
Y. Mao, Z. Zhao, M. Yang, W. Ding, and X.-P. Zhang are with Tsinghua-Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School, Tsinghua University, China. W. Ding is the corresponding author. E-mail: ({myz20, zhao-zh21,yml21} @mails.tsinghua.edu.cn, ding.wenbo@sz.tsinghua.edu.cn). W. Ding is also with RISC-V International Open Source Laboratory, Shenzhen, China, 518055.
L. Liang is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, and also with the Purple Mountain Laboratories, Nanjing 211111, China. E-mail: (lliang@seu.edu.cn)
Y. Liu is with the Institute for AI Industry Research (AIR), Tsinghua University, China. E-mail: (liuy03@air.tsinghua.edu.cn). Y. Liu and W. Ding are also with Shanghai AI Lab, Shanghai, China.
T. Lan is with the Department of Electrical and Computer Engineering, George Washington University, DC, USA. Email: (tlan@gwu.edu)
X.-P. Zhang is also with the Department of Electrical, Computer and Biomedical Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada. E-mail: (xzhang@ee.ryerson.ca)

updates on the fly. This compensation works even if sparse algorithms are employed, as we show that similarity properties are preserved under sparsity. We formally analyze the impact of such compensations in FL and prove that under bounded dissimilarity (i.e., the difference among sparse models produced by different clients are bounded) and a sufficiently small learning rate, the proposed SAFARI algorithm is guaranteed to converge. Extensive evaluations over several popular sparse algorithms (including MAG, Synflow [13] and FedSpa [14]) are conducted. The experiment results validate our theoretical analysis showing that the proposed SAFARI algorithm under unreliable communications achieves the same asymptotic convergence rate as vanilla FedAvg with reliable communications, even if 60% of the model weights are pruned and a large percentage (up to 80%) of client updates are lost in each round. SAFARI consistently achieves faster convergence than that without compensation under unreliable communications.

The contributions of this paper are summarized as follows.

- A sparsity-enabled robust FL framework, SAFARI, is proposed to simultaneously save communication overhead and cope with unreliable communications in FL, where the sparse algorithms are for transmitted model compression and a similarity-based compensation scheme is for bias reduction.
- We theoretically analyze the impact of such compensation with respect to sparse algorithms and prove that similarity properties are preserved under the use of sparse models. Besides, we establish global convergence analysis for SAFARI and demonstrate that even with limited and unreliable communications, SAFARI can achieve the same convergence rate of vanilla FedAvg with perfectly reliable communications.
- Experiments on the CIFAR-10 dataset validate our theoretical analysis, and SAFARI demonstrates fast and stable convergence under unreliable communications and outperforms baselines without compensation.

In summary, due to the engagement of sparse training and the robustness to unreliable communication, SAFARI works well with pruned models and lightweight communication protocol with no reliability guarantee. As a pioneer in tackling the potential aggregation bias resulting from unreliable communication through its measurement of local data distribution enabled by sparse transmission, SAFARI provides a systematic solution for communication-efficient federated learning under unreliable communication.

The rest of this paper is organized as follows. Section 2 introduces the background and related work as well as the motivation. In Section 3, the proposed method is described in detail. Theoretical analysis and the experimental results are provided in Section 4 and 5, respectively. Finally, concluding remarks are summarized in Section 6.

2 Background and Related work

2.1 Federated Learning

Assume a FL system with one central server and m distributed clients. Each client i in the client set $\mathbb{M} = \{1, \dots, m\}$ has a local dataset D_i of n_i data samples. The goal of federated training is to solve the following optimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{x}) = \sum_{i=1}^{m} w_i \mathcal{L}_i(\boldsymbol{x}), \tag{1}$$

where $\mathcal{L}_i(\boldsymbol{x}) = \sum_{i=1}^{n_i} \frac{1}{n_i} \sum_{z\in D_i} \ell_i(\boldsymbol{x}, z)$ is the local objective function at the i-th client. Specifically, z represents a data sample from D_i and $\ell_i : \mathbb{R}^d \to \mathbb{R}$ is the local loss function based on the learning model $\boldsymbol{x}$ and client i's own data. With $N = \sum_{i=1}^{m} n_i$, the aggregation weight $w_i = \frac{n_i}{N}$ for client i takes the local data size into consideration.

In the t-th communication round, the server first broadcasts the global model $\boldsymbol{x}^t$ to clients. Then each client independently runs τ local iterations by optimization solver such as the stochastic gradient descent (SGD) from the current global model $\boldsymbol{x}^t$ to optimize its own local objective function $\mathcal{L}_i(\boldsymbol{x})$. Take the SGD for example and the local iterations are as follows,

$$\begin{cases} \boldsymbol{x}_{i,0}^t = \boldsymbol{x}^t, \\ \boldsymbol{x}_{i,k}^t = \boldsymbol{x}_{i,k-1}^t - \eta g_i(\boldsymbol{x}_{i,k-1}^t | \xi_{i,k}), k \in \mathbb{K}, \end{cases} \tag{2}$$

where η is the learning rate, $g_i(\boldsymbol{x}_{i,k-1}^t|\xi_{i,k})$ is the stochastic gradient computed with the data batch $\xi_{i,k} \sim D_i$, $\boldsymbol{x}_{i,k}^t$ is the local model after k local iterations and $\mathbb{K} = \{1, \cdots, \tau\}$.

After completing τ iterations of local training, each client i will send the new model $\boldsymbol{x}_{i,\tau}^t$ back to the central server, and the server will aggregate the received client models to update the global model by:

$$\boldsymbol{x}^{t+1} = \sum_{i=1}^{m} w_i \boldsymbol{x}_{i,\tau}^t. \tag{3}$$

2.2 Practical Issues and Related Work

In the design and application of the FL system, there are some practical issues needed to be considered. According to a recent survey [1], the major issues in FL are summarized as convergence of SGD, aggregation robustness, upload frequency, privacy leakage and wireless connectivity. Among these five issues, robust aggregation, upload frequency, and wireless communications are all related to the capability and reliability of communications. In particular, the optimization of FL's global aggregation shown in (3) heavily relies on accurate, simultaneous, and up-to-date local updates received by the central server. Therefore, it deviates from the optimal with the presence of many realistic factors, such as local steps, updates compression, corruption and noise in communication links, and heterogeneous computational and communication capabilities for different local clients [1]. Given this strong relation between communication and robust aggregation, in this paper we mainly focus on strengthening aggregation robustness by addressing the bias caused by limited communication resources and unreliable communication links.

Limited Communication Resources. Edge devices in wireless networks usually have limited resources, especially for frequent communications. To reduce the transmission burden at each communication round, gradient or model weight compression is introduced, including quantization and sparsification. Gradient quantization maps each real-valued gradient/model element to a finite number of bits with lower-precision [15]–[17]. As another line of work, sparsification prunes the dense gradient/model with a large amount of non-zero elements to a sparser one. In practice, these two compression techniques can be jointly used, and sparsification is usually the first step to reduce the number of weights for further quantization and transmission. The simplest way to sparsify a model is to keep only the coordinates with large magnitudes exceeding a selected threshold [18]. More sophisticated methods like unbiased sparsification and variance-reduced sparsified SGD have also been developed for training in a distributed fashion [19]–[22]. One remaining question is that such sparsification operates after the local training completes, which provides no reduction on the computation and memory cost during training.

As the training model becomes larger along with the growth of training data in recent years, sparse learning that pre-conducts sparsification and maintains sparse structure throughout training has been intensively investigated. In [23], fully-connected layers were replaced by sparse ones achieved from an initial sparse topology with evolutionary algorithms before training. The connection sensitivity has been investigated in [24] for Single-Shot Network Pruning (SNIP). In [25], the exponentially smoothed gradients were utilized to identify model layers and weights which reduced the error efficiently. You *et al.* proposed to use the change of mask distances between epochs to identify a small sub-network at the early training stage, which could restore the comparable test accuracy to the dense network when being trained independently [26]. Moreover, the sparse topology's updates based on parameter magnitudes and infrequent gradient calculations in [27] loosened the limitation on the size relationship between the sparse model and the corresponding dense model, which further reduced the computation cost for sparse learning. However, despite the success empirical performance of the above sparse learning methods, theoretical analysis of the sparse model's property is still limited.

Unreliable Communications and Local Bias. Due to the limited capability of distributed clients and communication channels, communication reliability cannot be guaranteed in the FL system, especially with wireless networks [1]. A previous work has proposed to address the unreliable issues by optimizing the aggregation weights according to the link reliability matrix of communication links in a decentralized network [11]. Thus, it requires the knowledge of reliability matrix in advance, as presented in the following *independent and stable links* assumption from [11]:

- **Independent and stable links.** *The packet transmission on different links are independent and the link reliability matrix* $\mathbf{P}$ *remains fixed during training. Specifically,* $\mathbf{P} = [p_{i,j}] \in \mathbb{R}^{m \times m}$ *describes the level of link reliability in the communication network, where* $p_{i,j}$ *represents the probability of successful transmission from the* i*-th device to the* j*-th device and* $p_{i,i} = 0, \forall i \in \{1, \ldots, m\}$.

In the above assumption, the unreliability characteristics measured by the probability of successful transmission for each link in the communication network is fixed and known in advance. However, the link reliability matrix is sometimes infeasible in real-world systems.

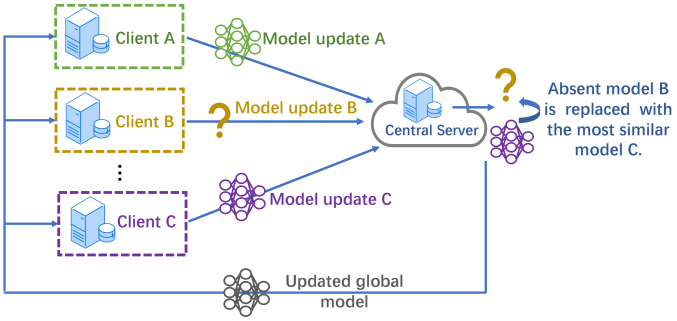

Fig. 1: The schematic illustration of SAFARI.

To tackle the bias caused by local training steps, methods like drift-reduced SCAFFOLD [28] and Inexact DANE [29] with local approximate sub-solver have been developed and shown to be effective when the heterogeneity of local objectives is small enough. Recently, the Bias-Variance Reduced Local SGD algorithm surpasses non-local methods under a more relaxed second-order heterogeneity assumption [30]. But the existing bias-reduction techniques still rely on reliable communications that guarantee the successful transmission of local updates.

2.3 Motivating Applications

In this section, we provide several examples to explain some useful properties in practical FL systems that can be utilized to address the aforementioned issues.

Local Computing Resources. In FL collaborative systems, local clients are always equipped with a certain degree of computing power, even for small edge devices as smartphones, werables and sensors, or distributed medical/financial institutes. It makes local sparse learning feasible, and reveals great potentials to achieve highly efficient local training with limited distributed resources.

Clusterable Clients. Although the non-IID data distribution and unstable communication remain challenging in FL systems, it is noted that the clients in quite a few real-word systems tend to be clusterable in terms of data distribution. For example, in an Internet of vehicles system, vehicles within a certain area tend to record similar transportation information. Besides, the devices within the same smart home system usually collect the features of the same person. In these examples, the dissimilarity between client data in a certain group may be small, or even better, they follow IID data distribution for the same learning task.

Note that although the pervasiveness of clusterable clients is demonstrated, the following analysis of our method is built upon the standard assumption on data dissimilarity as previous works [31].

3 Methods

In this section, we introduce our approach and elaborate on details of the proposed SAFARI algorithm as illustrated in Figure 1.

3.1 Core Concepts and Approach

Here we first describe the two building blocks of the proposed SAFARI algorithm, which are *the sparsity-enabled communication efficiency* and *the similarity assisted bias reduction with unreliable wireless communications*. The target problem and the proposed solution are explained in detail.

FL with Limited and Unreliable Communications: According to the previous work, the lightweight message based connectionless protocol UDP is commonly used in resource-limited wireless communications. Specifically, UDP reduces much overhead by omitting mechanisms like ACK message confirmation and lost package retransmission [11]. In exchange for the relatively low communication overhead, the transmission reliability can not be guaranteed in UDP transmissions. Assume a link reliability list $P = \{p_1^t, \ldots, p_m^t\}$, where $0 \leq p_i^t \leq 1$ is the probability that the server successfully receives the local model $\boldsymbol{x}_{i,\tau}^t$ from client i at the communication round t. In real-world scenarios, each server-client link's reliability could depend on several factors, i.e., the quality of the channel, the distance between the central server and the corresponding client, as well as the reliability of the client device.

Sparsity-enabled Communication Efficiency and Similarity assisted Bias Reduction: To save computing resources and training/inference time, sparse learning on large neural networks has been widely deployed in the deep learning field [7], [23], [24], [27], [32]. When being introduced to FL scenarios, it can save the communication overload by reducing the amount of model weights to be sent. In this context, we propose to conduct the sparse learning at local clients, and utilize the similarity of sparse models to address the bias caused by unreliable communications. Concretely, the server keeps a record of the similarity across clients, which is measured by the sparse models they produce. The similarity record changes along with the training process according to the sparse models successfully received at each global round. With this record, for inactive clients whose models have not been received by the server (client fails to participate in training or encounters network failure), the missing model is substituted by the model from the most similar active client.

We will show in the theoretical part that in such way, the bias caused by random loss of local updates can be entirely eliminated when the clients are clusterable, or at least limited to the same order of the intrinsic data dissimilarity bound in more general scenarios. This enables us to keep the same asymptotic convergence rate as vanilla FedAvg with perfectly reliable communications.

3.2 The SAFARI Algorithm

The proposed SAFARI algorithm to address the limited and unreliable communication issue is summarized in Algorithm 1. As in vanilla FedAvg [2], the server first initializes an original global model $\boldsymbol{x}^0$ and broadcasts it through communication links. Due to the unreliability of communications, some clients may fail to receive the global model from the server. For each client i that successfully receives the global model, it performs local sparse training as illustrated in Algorithm 2. Specifically, it first calculates a mask $\mathcal{M}_i$ based on a specific sparse algorithm to sparsify the global model's structure, and then performs local SGD with the sparse structure for τ iterations. Once the local sparse training is completed, the client will send the sparse local model $\boldsymbol{x}_{i,\tau}^t$ back to the server.

Algorithm 1 SAFARI

```
Input: The number of communication rounds $T$, the learning rate $\eta$, the number of local steps $\tau$.
Initialize: The initial dense global model $\boldsymbol{x}^0$.
  for $t = 0$ to $T - 1$ do
    Server broadcasts $\boldsymbol{x}^t$ to all clients.
    for each client $i$ receives the message in parallel do
      Perform Local Sparse Training($\boldsymbol{x}^t, \eta, \tau$).
      Send the updated sparse model $\boldsymbol{x}_{i,\tau}^t$ back to the server.
    end for
    Server performs Bias Reduced Global Aggregation.
  end for
  Finish the training with global model $\boldsymbol{x}^T$.
```

Algorithm 2 Local Sparse Training.

```
Input: The received global model $\boldsymbol{x}^t$, the learing rate $\eta$, the number of local steps $\tau$.
  Calculate mask $\mathcal{M}_i$ based on a specific sparse algorithm.
  Prune the model for a sparser structure: $\boldsymbol{x}_{i,0}^t = \boldsymbol{x}^t \odot \mathcal{M}_i$.
  for $k = 1$ to $\tau$ do
    Sample a mini-batch $\xi_{i,k}$ from local dataset $D_i$.
    Compute the local gradient $g_i(\boldsymbol{x}_{i,0}^t|\xi_{i,k})$.
    Local SGD step: $\boldsymbol{x}_{i,k}^t = \boldsymbol{x}_{i,k-1}^t - \eta g_i(\boldsymbol{x}_{i,k-1}^t|\xi_{i,k})$.
  end for
  return $\boldsymbol{x}_{i,\tau}^t$.
```

Again, since the communications are unreliable, not all of the updated local models can be received by the server. The proposed global aggregation with similarity-based compensation is summarized as Algorithm 3. To address the potential bias caused by such random loss of client updates, the server will determine the active client group $\mathbb{M}_a$ based on the received client models. Before the aggregation, the server will update the similarity matrix among active clients, and then replace the model from each missing client j with the received model from the most similar active client j'. After the total T global rounds, the FL training is completed with a trained global model $\boldsymbol{x}^T$.

3.3 Comparison with Previous Works

In this section, we summarize the difference between the proposed SAFARI compared with representative previous methods and highlight its contributions in Table 1. FedAvg [2] invented the idea of FL with the key idea of reducing the communication costs required for global convergence in distributed systems. More advanced methods like SCAFFOLD [28] and FedProx [33] took a step further to consider not only reducing communication costs but also addressing data heterogeneity in FL systems. To further save communication resources, the representative STC [22] developed a sparsification technique that stays robust to

TABLE 1. Algorithm comparison regarding practical factors considered

Method	Communication Cost	Computation/Memory Cost	Data Heterogeneity	Sparsity	Unreliable Communication
FedAvg [2]	✓	×	×	×	×
SCAFFOLD [28]	✓	×	✓	×	×
FedProx [33]	✓	×	✓	×	×
STC [22]	✓	×	✓	✓	×
FedSpa [14]	✓	✓	✓	✓	×
SAFARI	✓	✓	✓	✓	✓

data heterogeneity. Then, also inspired by sparse learning methods for centralized learning, methods like FedSpa [14] introduced sparse learning methods to the FL regime, which further achieved computation/memory costs based on previous FL methods.

In this work, the SAFARI framework considers all the factors involved in the aforementioned works. More specifically, it applies sparse learning to save communication/computation/memory costs simultaneously, and it utilizes the sparse model similarity to measure the heterogeneity in underlying data distribution among clients.

Furthermore, SAFARI also explores how to address the unreliable communication issues in FL systems. By substituting missing model updates with the most similar received model updates in each communication, SAFARI is a pioneer in tackling the potential aggregation bias resulting from unreliable communication through its measurement of local data distribution enabled by sparsified transmission. Generally speaking, the proposed SAFARI framework is a systematic solution considering multiple key factors for practical FL applications.

Algorithm 3 Global Aggregation with Similarity-based Compensation.

```
Input: The received client models, the whole client set $\mathbb{M}$, the active client group $\mathbb{M}_a \neq \varnothing$, and $s$ similarity function (e.g. Euclidean Distance).
  for each client $i$ whose model has been received do
    $\mathbb{M}_a = \mathbb{M}_a \cup \{i\}$.
  end for
  Server updates the similarity matrix $\rho \in \mathbb{R}^{m \times m}$ with $\rho_{u,v} = s(\boldsymbol{x}_{u,\tau}^t, \boldsymbol{x}_{v,\tau}^t), \forall u, v \in \mathbb{M}_a, u \neq v$.
  for each client $j \in \mathbb{M} \setminus \mathbb{M}_a$ do
    $j' \leftarrow i \in \mathbb{M}_a$ that maximizes $\rho_{i,j}$.
  end for
  Server performs global aggregation:
  $\boldsymbol{x}^{t+1} = \sum_{i \in \mathbb{M}_a} w_i \boldsymbol{x}_{i,\tau}^t + \sum_{j \in \mathbb{M} \setminus \mathbb{M}_a} w_j \boldsymbol{x}_{j',\tau}^t$.
  return $\boldsymbol{x}^{t+1}$.
```

4 Theoretical Analysis

In this section, we analyze the convergence property of our method and theoretically prove that it can achieve the same convergence rate as the vanilla FedAvg with reliable communications [31].

Notation. In the following part, we use $\|\boldsymbol{x}\|$, $\|\boldsymbol{x}\|_1$ and $[\boldsymbol{x}]_n$ to denote the l_2, l_1 norms and the n-th element of a vector $\boldsymbol{x}$, respectively.

4.1 Assumptions

4.1.1 Functions

We first adopt the following three standard assumptions on functions, which are widely used in analyzing convergence behavior of non-convex optimization problems:

- **Smoothness.** *The local objective functions are L-smooth, i.e.,* $\forall i \in \mathbb{M}$:
$$\|\nabla \mathcal{L}_i(\boldsymbol{x}) - \nabla \mathcal{L}_i(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d. \quad (4)$$
- **Unbiased Gradient and Bounded Variance.** $\forall i \in \mathbb{M}$, *the stochastic gradient* $g_i(\boldsymbol{x}|\xi)$ *calculated with local data batch* ξ *is an unbiased estimator of the local gradient:* $\mathbb{E}_{\xi \sim D_i}[g_i(\boldsymbol{x}|\xi)] = \nabla \mathcal{L}_i(\boldsymbol{x})$, *and the variance is bounded by:* $\mathbb{E}_{\xi \sim D_i}\|g_i(\boldsymbol{x}|\xi) - \nabla \mathcal{L}_i(\boldsymbol{x})\|^2 \leq \sigma^2$, $\forall \boldsymbol{x} \in \mathbb{R}^d$, $\sigma^2 \geq 0$.
- **Bounded Dissimilarity.** *There exist constants* $\beta^2 \geq 1$ *and* $\zeta^2 \geq 0$ *such that*:
$$\sum_{i=1}^{m} w_i \|\nabla \mathcal{L}_i(\boldsymbol{x})\|^2 \leq \beta^2 \left\| \sum_{i=1}^{m} w_i \nabla \mathcal{L}_i(\boldsymbol{x}) \right\|^2 + \zeta^2. \quad (5)$$
Particularly, $\beta^2 = 1$ *and* $\zeta^2 = 0$ *indicate the IID situation where all the local functions are identical.*

4.1.2 Sparse Models

To analyze the property of local training with sparse models, a common assumption on the mask-induced error is also adopted from sparsification-related literature [7].

- **Mask-induced Error.** *It is assumed that* $\forall \boldsymbol{x} \in \mathbb{R}^d$, *the corresponding binary mask* $\mathcal{M} \in \{0,1\}^d$ *satisfy*
$$\|\boldsymbol{x} \odot \mathcal{M} - \boldsymbol{x}\|^2 \leq \delta^2 \|\boldsymbol{x}\|^2, 0 < \delta < 1, \quad (6)$$
where $\odot$ *denotes the Hadamard product.*

Note that the above assumption is quite a relaxed one, which is not limited to any specific sparse algorithms. Furthermore, to analyze the impact of sparse learning in distributed fashion, we make an assumption on the similarity between local training with sparse structures.

- **Similarity Preservation.** *Under the bounded dissimilarity assumption,* $\forall \boldsymbol{x} \in \mathbb{R}^d$, $\forall i, j \in \mathbb{M}$ *and local model mask* $\{\mathcal{M}_i\}_{i=1}^m$:
$$\|\nabla \mathcal{L}_i(\boldsymbol{x} \odot \mathcal{M}_i)\|^2 \leq \beta^2 \|\nabla \mathcal{L}_j(\boldsymbol{x} \odot \mathcal{M}_j)\|^2 + \zeta^2. \quad (7)$$

The above assumption indicates the rationality behind the compensation based on the similarity among sparse models produced by different clients. In this paper, we have analyzed the establishment of this assumption with regard to SNIP sparse algorithm [24] in appendix A, and empirically shown that this assumption will hold for most existing sparse algorithms..

4.1.3 Communication Networks

Similar to [11], we also make an additional assumption on the unreliable communication network. But compared with the *independent and stable links* assumption made therein, we extend the condition to cover *independent and unstable links*. In other words, the algorithm proposed in this paper does not require the link reliability to be known in advance or keep stable during training.

- **Independent and Unstable Links.** *The transmissions on different client links are independent and each link's reliability may change during training process.*

4.2 Descent Lemma with Sparsification

Lemma 1. (Descent Lemma with Sparsification) *With the above assumption on function smoothness, unbiased gradient and bounded variance, as well as sparsification, if $\eta \le \tau/(6L)$, it holds $\forall i \in \mathbb{M}$, $t \in \mathbb{T} = \{0, \dots, T-1\}$, $k \in \mathbb{K}$ that,*

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}_i(\boldsymbol{x}_{i,k}^t)\right] \le & \,\mathbb{E}\left[\mathcal{L}_i(\boldsymbol{x}_{i,k-1}^t)\right] - \frac{\eta}{3\tau}\mathbb{E}\left\|\nabla\mathcal{L}_i(\boldsymbol{x}_{i,k-1}^t)\right\|^2 \\ & + \frac{\eta^2 L\sigma^2}{2\tau^2} + \frac{2\eta L^2\delta^2}{3\tau}\mathbb{E}\left\|\boldsymbol{x}_{i,k-1}^t\right\|^2. \end{aligned} \tag{8}$$

We refer the readers to appendix A for proof details. From Lemma 1, with the appropriate learning rate, the local objective value will decrease by $\frac{\eta}{3\tau}\mathbb{E}\|\nabla\mathcal{L}_i(\boldsymbol{x}_{i,k-1}^t)\|^2$ after every local step. The lemma also meets the expectation that the training will suffer from stochastic gradient variance σ and weight pruning error δ. To the best of the authors' knowledge, rigorous analysis to quantify the weight pruning error δ is still lacking and also beyond the scope of this work. That said, empirical success of popular sparse algorithms implies that this error is quite tolerable in practice [7], [23], [24], [27], [32], which enables us to implement sparse training in FL for communication efficiency, and meanwhile utilizes the properties of sparse models to address the unreliable communications.

4.3 Global Convergence

To keep consistent and fair comparison with existing FL researches, we build our analysis within the general analysis framework for heterogeneous federated optimization algorithms proposed by [31]. Similarly, we first quantify the model update between rounds. From the global point of view, $\boldsymbol{x}_{i,\tau}^t$ represents the local model sent to server after client i's local iterations, which is supposed to be a sparse one. Recall that the global model is updated by the following rule under reliable communications:

$$\boldsymbol{x}^{t+1} = \sum_{i=1}^m w_i \boldsymbol{x}_{i,\tau}^t = \boldsymbol{x}^t - \tau\eta\sum_{i=1}^m w_i \boldsymbol{d}_i^{(t)}, \tag{9}$$

where $\boldsymbol{d}_i^{(t)} = \frac{1}{\tau}\sum_{k=1}^{\tau} g_i(\boldsymbol{x}_{i,k}^t)$ is the normalized stochastic gradient at client i. Correspondingly, the normalized gradient at each client is defined as

$$\boldsymbol{h}_i^{(t)} = \frac{1}{\tau}\sum_{k=1}^{\tau}\nabla\mathcal{L}_i(\boldsymbol{x}_{i,k}^t), i \in \mathbb{M}. \tag{10}$$

To solve the problem caused by unreliable communications, the global model is updated with the proposed compensation based on sparse model similarity. Therefore, the expectation of global model update can be written as

$$\mathbb{E}\left[\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\right] = -\tau\eta\sum_{i=1}^m w_i\left[p_i^t \boldsymbol{d}_i^{(t)} + \left(1-p_i^t\right)\boldsymbol{d}_{i'}^{(t)}\right], \tag{11}$$

where i' is the index of the most similar client used for replacing client i in case it is lost, and p_i^t is the reliability of the channel between client i and the server at round t.

According to the smoothness assumption, there is,

$$\begin{aligned} & \mathbb{E}\left[\mathcal{L}(\boldsymbol{x}^{t+1})\right] - \mathcal{L}(\boldsymbol{x}^t) \\ \le & \,\mathbb{E}\left\langle\nabla\mathcal{L}(\boldsymbol{x}^t), \boldsymbol{x}^{t+1} - \boldsymbol{x}^t\right\rangle + \frac{L}{2}\mathbb{E}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2 \qquad (12) \\ = & -\tau\eta\,\underbrace{\mathbb{E}\left\langle\nabla\mathcal{L}(\boldsymbol{x}^t), \sum_{i=1}^m w_i\left[p_i^t \boldsymbol{d}_i^{(t)} + \left(1-p_i^t\right)\boldsymbol{d}_{i'}^{(t)}\right]\right\rangle}_{T_1} \\ & + \frac{\tau^2\eta^2 L}{2}\,\underbrace{\mathbb{E}\left\|\sum_{i=1}^m w_i\left[p_i^t \boldsymbol{d}_i^{(t)} + \left(1-p_i^t\right)\boldsymbol{d}_{i'}^{(t)}\right]\right\|^2}_{T_2}. \qquad (13) \end{aligned}$$

Based on the above results, the following Lemma 2 and Lemma 3 provide a milestone for analyzing the convergence property of the update rule (11) by bounding the T_1 and T_2 terms in (13).

Lemma 2. *With the above assumptions, if $\eta \le \frac{1}{3\tau L}$, the left hand side of* (13) *can be bounded as follows,*

$$\begin{aligned} & \frac{1}{\tau\eta}\left(\mathbb{E}\left[\mathcal{L}(\boldsymbol{x}^{t+1})\right] - \mathcal{L}(\boldsymbol{x}^t)\right) \\ \le & -\left\|\nabla\mathcal{L}(\boldsymbol{x}^t)\right\|^2 + \tau\eta L\sum_{i=1}^m w_i^2\left[6 + 9\left(1-p_i^t\right)^2\right]\sigma^2 \\ & + \left(\frac{3}{2}\tau\eta L - \frac{1}{2}\right)\sum_{i=1}^m w_i^2\left(1-p_i^t\right)^2\mathbb{E}\|\boldsymbol{h}_{i'}^{(t)} - \boldsymbol{h}_i^{(t)}\|^2 \\ & + \frac{1}{2}\sum_{i=1}^m w_i\mathbb{E}\|\nabla\mathcal{L}_i(\boldsymbol{x}^t) - \boldsymbol{h}_i^{(t)}\|^2, \end{aligned} \tag{14}$$

Proof. For the first term on the right hand side of (13),

$$\begin{aligned} T_1 = & \,\mathbb{E}\left\langle\nabla\mathcal{L}(\boldsymbol{x}^t), \sum_{i=1}^m w_i p_i^t(\boldsymbol{d}_i^{(t)} - \boldsymbol{h}_i^{(t)} + \boldsymbol{h}_i^{(t)})\right\rangle \\ & + \mathbb{E}\left\langle\nabla\mathcal{L}(\boldsymbol{x}^t), \sum_{i=1}^m w_i\left(1-p_i^t\right)(\boldsymbol{d}_{i'}^{(t)} - \boldsymbol{h}_{i'}^{(t)} + \boldsymbol{h}_{i'}^{(t)})\right\rangle \\ = & \,\mathbb{E}\left\langle\nabla\mathcal{L}(\boldsymbol{x}^t), \sum_{i=1}^m w_i p_i^t \boldsymbol{h}_i^{(t)}\right\rangle \\ & + \mathbb{E}\left\langle\nabla\mathcal{L}(\boldsymbol{x}^t), \sum_{i=1}^m w_i\left(1-p_i^t\right)\boldsymbol{h}_{i'}^{(t)}\right\rangle, \end{aligned} \tag{15}$$

where the second equality comes from the unbiased gradient assumption which implies $\mathbb{E}[\boldsymbol{d}_i^{(t)} - \boldsymbol{h}_i^{(t)}] = 0$. With some arrangements, the T_1 term can be written as,

$$\begin{aligned} T_1 = & \,\mathbb{E}\left\langle\nabla\mathcal{L}(\boldsymbol{x}^t), \sum_{i=1}^m w_i \boldsymbol{h}_i^{(t)}\right\rangle \qquad (16) \\ & + \mathbb{E}\left\langle\nabla\mathcal{L}(\boldsymbol{x}^t), \sum_{i=1}^m w_i\left(1-p_i^t\right)(\boldsymbol{h}_{i'}^{(t)} - \boldsymbol{h}_i^{(t)})\right\rangle \end{aligned}$$

$$
\begin{aligned}
\leq&\frac{1}{2}\left\|\nabla\mathcal{L}(\boldsymbol{x}^t)\right\|^2+\frac{1}{2}\mathbb{E}\left\|\sum_{i=1}^{m}w_i\boldsymbol{h}_i^{(t)}\right\|^2\\
&+\frac{1}{2}\left\|\nabla\mathcal{L}(\boldsymbol{x}^t)\right\|^2+\frac{1}{2}\mathbb{E}\left\|\sum_{i=1}^{m}w_i\left(1-p_i^t\right)\left(\boldsymbol{h}_{i'}^{(t)}-\boldsymbol{h}_i^{(t)}\right)\right\|^2\\
&-\frac{1}{2}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{x}^t)-\sum_{i=1}^{m}w_i\boldsymbol{h}_i^{(t)}\right\|^2,
\end{aligned}
\tag{17}
$$

where the last inequality follows from $2\langle\boldsymbol{a},\boldsymbol{b}\rangle=\|\boldsymbol{a}\|^2+\|\boldsymbol{b}\|^2-\|\boldsymbol{a}-\boldsymbol{b}\|^2$.

For the second term on the right hand side of(13), with $\|\sum_{i=1}^{n}\boldsymbol{w}_i\|_2^2\leq n\sum_{i=1}^{n}\|\boldsymbol{w}_i\|_2^2$, there is,

$$
\begin{aligned}
T_2=&\,\mathbb{E}\left\|\sum_{i=1}^{m}w_i\boldsymbol{d}_i^{(t)}+\sum_{i=1}^{m}w_i\left[\left(1-p_i^t\right)\left(\boldsymbol{d}_{i'}^{(t)}-\boldsymbol{d}_i^{(t)}\right)\right]\right\|^2\\
=&\,\mathbb{E}\left\|\sum_{i=1}^{m}w_i\boldsymbol{h}_i^{(t)}+\sum_{i=1}^{m}w_i(\boldsymbol{d}_i^{(t)}-\boldsymbol{h}_i^{(t)})+\sum_{i=1}^{m}w_i\left[\left(1-p_i^t\right)\left(\boldsymbol{d}_{i'}^{(t)}-\boldsymbol{d}_i^{(t)}\right)\right]\right\|^2\\
\leq&\,3\mathbb{E}\left\|\sum_{i=1}^{m}w_i\boldsymbol{h}_i^{(t)}\right\|^2+12\mathbb{E}\left\|\sum_{i=1}^{m}w_i(\boldsymbol{d}_i^{(t)}-\boldsymbol{h}_i^{(t)})\right\|^2\\
&+3\mathbb{E}\left\|\sum_{i=1}^{m}w_i\left(1-p_i^t\right)\left(\boldsymbol{h}_{i'}^{(t)}-\boldsymbol{h}_i^{(t)}\right)\right\|^2\\
&+12\mathbb{E}\left\|\sum_{i=1}^{m}w_i\left(1-p_i^t\right)\left(\boldsymbol{h}_i^{(t)}-\boldsymbol{d}_i^{(t)}\right)\right\|^2\\
&+6\mathbb{E}\left\|\sum_{i=1}^{m}w_i\left(1-p_i^t\right)\left(\boldsymbol{d}_{i'}^{(t)}-\boldsymbol{h}_{i'}^{(t)}\right)\right\|^2
\end{aligned}
\tag{18}
$$

$$
\begin{aligned}
=&\,3\mathbb{E}\left\|\sum_{i=1}^{m}w_i\boldsymbol{h}_i^{(t)}\right\|^2+12\sum_{i=1}^{m}w_i^2\mathbb{E}\|\boldsymbol{d}_i^{(t)}-\boldsymbol{h}_i^{(t)}\|^2\\
&+3\sum_{i=1}^{m}w_i^2\left(1-p_i^t\right)^2\mathbb{E}\|\boldsymbol{h}_{i'}^{(t)}-\boldsymbol{h}_i^{(t)}\|^2\\
&+12\sum_{i=1}^{m}w_i^2\left(1-p_i^t\right)^2\mathbb{E}\|\boldsymbol{h}_i^{(t)}-\boldsymbol{d}_i^{(t)}\|^2\\
&+6\sum_{i=1}^{m}w_i^2\left(1-p_i^t\right)^2\mathbb{E}\|\boldsymbol{d}_{i'}^{(t)}-\boldsymbol{h}_{i'}^{(t)}\|^2.
\end{aligned}
\tag{19}
$$

With the assumption on the gradient variance, the second term can then be bounded as,

$$
\begin{aligned}
T_2\leq&\sum_{i=1}^{m}w_i^2\left[12+18\left(1-p_i^t\right)^2\right]\sigma^2+3\mathbb{E}\left\|\sum_{i=1}^{m}w_i\boldsymbol{h}_i^{(t)}\right\|^2\\
&+3\sum_{i=1}^{m}w_i^2\left(1-p_i^t\right)^2\mathbb{E}\|\boldsymbol{h}_i^{(t)}-\boldsymbol{h}_{i'}^{(t)}\|^2.
\end{aligned}
\tag{20}
$$

Plugging the bound on T_1 (17) and the bound on T_2 (20) back into (13), there is,

$$
\begin{aligned}
&\mathbb{E}\left[\mathcal{L}(\boldsymbol{x}^{t+1})\right]-\mathcal{L}(\boldsymbol{x}^t)\\
\leq&-\tau\eta\left\|\nabla\mathcal{L}(\boldsymbol{x}^t)\right\|^2+\left(\frac{3}{2}\tau^2\eta^2L-\frac{\tau\eta}{2}\right)\mathbb{E}\left\|\sum_{i=1}^{m}w_i\boldsymbol{h}_i^{(t)}\right\|^2\\
&+\frac{\tau\eta}{2}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{x}^t)-\sum_{i=1}^{m}w_i\boldsymbol{h}_i^{(t)}\right\|^2\\
&+\left(\frac{3}{2}\tau^2\eta^2L-\frac{\tau\eta}{2}\right)\sum_{i=1}^{m}w_i^2\left(1-p_i^t\right)^2\mathbb{E}\|\boldsymbol{h}_{i'}^{(t)}-\boldsymbol{h}_i^{(t)}\|^2\\
&+\tau^2\eta^2L\sum_{i=1}^{m}w_i^2\left[6+9\left(1-p_i^t\right)^2\right]\sigma^2.
\end{aligned}
\tag{21}
$$

If $\tau\eta L\leq\frac{1}{3}$, it holds that,

$$
\begin{aligned}
&\frac{1}{\tau\eta}\left(\mathbb{E}\left[\mathcal{L}(\boldsymbol{x}^{t+1})\right]-\mathcal{L}(\boldsymbol{x}^t)\right)\\
\leq&-\left\|\nabla\mathcal{L}(\boldsymbol{x}^t)\right\|^2+\tau\eta L\sum_{i=1}^{m}w_i^2\left[6+9\left(1-p_i^t\right)^2\right]\sigma^2\\
&+\left(\frac{3}{2}\tau\eta L-\frac{1}{2}\right)\sum_{i=1}^{m}w_i^2\left(1-p_i^t\right)^2\mathbb{E}\|\boldsymbol{h}_{i'}^{(t)}-\boldsymbol{h}_i^{(t)}\|^2\\
&+\frac{1}{2}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{x}^t)-\sum_{i=1}^{m}w_i\boldsymbol{h}_i^{(t)}\right\|^2
\end{aligned}
\tag{22}
$$

$$
\begin{aligned}
\leq&-\left\|\nabla\mathcal{L}(\boldsymbol{x}^t)\right\|^2+\tau\eta L\sum_{i=1}^{m}w_i^2\left[6+9\left(1-p_i^t\right)^2\right]\sigma^2\\
&+\left(\frac{3}{2}\tau\eta L-\frac{1}{2}\right)\sum_{i=1}^{m}w_i^2\left(1-p_i^t\right)^2\mathbb{E}\|\boldsymbol{h}_{i'}^{(t)}-\boldsymbol{h}_i^{(t)}\|^2\\
&+\frac{1}{2}\sum_{i=1}^{m}w_i\mathbb{E}\|\nabla\mathcal{L}_i(\boldsymbol{x}^t)-\boldsymbol{h}_i^{(t)}\|^2,
\end{aligned}
\tag{23}
$$

where the last inequality comes from Jensen's Inequality $\left\|\sum_{i=1}^{m}w_i\boldsymbol{a}_i\right\|^2\leq\sum_{i=1}^{m}w_i\left\|\boldsymbol{a}_i\right\|^2$. The proof of Lemma 2 is completed.

Lemma 3. *With the above assumptions, if $\eta\leq\frac{1}{2\tau L}$, the difference between the gradient computed with global model and normalized client gradient can be bounded as follows,*

$$
\begin{aligned}
&\sum_{i=1}^{m}w_i\mathbb{E}\|\nabla\mathcal{L}_i(\boldsymbol{x}^t)-\boldsymbol{h}_i^{(t)}\|^2\\
&\leq\frac{2\eta^2\sigma^2L^2\left(\tau-1\right)}{1-\gamma}+\frac{\gamma\beta^2}{1-\gamma}\mathbb{E}\|\nabla\mathcal{L}(\boldsymbol{x}^t)\|^2+\frac{\gamma\zeta^2}{1-\gamma},
\end{aligned}
\tag{24}
$$

where $\gamma=4\eta^2L^2\tau\left(\tau-1\right)$. Since the compensation strategy for unreliable channels is not involved in the conclusion of this Lemma, we refer the readers to appendix A for proof details.

Global Convergence Property. The following theorem indicates the global convergence property of the proposed method with unreliable communications based on Lemma2 and Lemma 3.

Theorem 1 *Under the above assumptions, if $\eta\leq\frac{1}{3\tau L}$, the optimization error after total T iterations is bounded as follows*:

$$
\begin{aligned}
\min_{t\in\mathbb{T}}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{x}^t)\right\|^2\leq&\,\mathcal{O}\left(\frac{1}{\sqrt{m\tau T}}\right)+\mathcal{O}\left(\frac{A\sigma^2}{\sqrt{m\tau T}}\right)\\
&+\mathcal{O}\left(\frac{mB\sigma^2}{\tau T}\right)+\mathcal{O}\left(\frac{mC\zeta^2}{\tau T}\right),
\end{aligned}
\tag{25}
$$

where $A=\tau,B=\tau-1,C=\tau(\tau-1)$, and all other constants are subsumed in $\mathcal{O}$.

Proof. Combining the Lemma 2 conclusion (23) and Lemma 3 conclusion (24) together, we can then bound the objective reduction in this way,

$$
\frac{1}{\tau\eta}\left(\mathbb{E}\left[\mathcal{L}(\boldsymbol{x}^{t+1})\right]-\mathcal{L}(\boldsymbol{x}^t)\right)
$$

$$\begin{aligned}
&\leq -\|\nabla \mathcal{L}(\boldsymbol{x}^t)\|^2 + \tau\eta L \sum_{i=1}^{m} w_i^2 \left[6 + 9\left(1 - p_i^t\right)^2\right]\sigma^2 \\
&+ \frac{\gamma\beta^2}{2(1-\gamma)}\mathbb{E}\|\nabla \mathcal{L}(\boldsymbol{x}^t)\|^2 + \frac{\gamma\zeta^2}{2(1-\gamma)} \\
&+ \left(\frac{3}{2}\tau\eta L - \frac{1}{2}\right)\sum_{i=1}^{m} w_i^2 \left(1 - p_i^t\right)^2 \mathbb{E}\|\boldsymbol{h}_{i'}^{(t)} - \boldsymbol{h}_i^{(t)}\|^2 \\
&+ \frac{\eta^2\sigma^2L^2\left(\tau - 1\right)}{1-\gamma}.
\end{aligned} \tag{26}$$

Due to the bounded dissimilarity assumption on sparse models (7), we have

$$\begin{aligned}
&\frac{1}{\tau\eta}\left(\mathbb{E}\left[\mathcal{L}(\boldsymbol{x}^{t+1})\right] - \mathcal{L}(\boldsymbol{x}^t)\right) \\
&\leq -\|\nabla \mathcal{L}(\boldsymbol{x}^t)\|^2 + \tau\eta L \sum_{i=1}^{m} w_i^2 \left[6 + 9\left(1 - p_i^t\right)^2\right]\sigma^2 \\
&+ 2\left(\frac{3}{2}\tau\eta L - \frac{1}{2}\right)\left[\sum_{i=1}^{m} w_i^2 \mathbb{E}\|\boldsymbol{h}_i^{(t)}\|^2 + \sum_{i=1'}^{m} w_i^2 \mathbb{E}\|\boldsymbol{h}_{i'}^{(t)}\|^2\right] \\
&+ \frac{\eta^2\sigma^2L^2\left(\tau - 1\right)}{1-\gamma} + \frac{\gamma\beta^2}{2(1-\gamma)}\mathbb{E}\|\nabla \mathcal{L}(\boldsymbol{x}^t)\|^2 + \frac{\gamma\zeta^2}{2(1-\gamma)}. \\
&\leq -\|\nabla \mathcal{L}(\boldsymbol{x}^t)\|^2 + \tau\eta L \sum_{i=1}^{m} w_i^2 \left[6 + 9\left(1 - p_i^t\right)^2\right]\sigma^2 \\
&+ 4\left(\frac{3}{2}\tau\eta L - \frac{1}{2}\right)\beta^2 \sum_{i=1}^{m} w_i^2 \mathbb{E}\|\nabla \mathcal{L}(\boldsymbol{x}^t)\|^2 \\
&+ 4\left(\frac{3}{2}\tau\eta L - \frac{1}{2}\right)\sum_{i=1}^{m} w_i^2 \zeta^2 + \frac{\gamma\zeta^2}{2(1-\gamma)} \\
&+ \frac{\eta^2\sigma^2L^2\left(\tau - 1\right)}{1-\gamma} + \frac{\gamma\beta^2}{2(1-\gamma)}\mathbb{E}\|\nabla \mathcal{L}(\boldsymbol{x}^t)\|^2.
\end{aligned} \tag{27}$$

If $\eta \leq \frac{1}{3\tau L}$, it holds that,

$$\begin{aligned}
&\frac{1}{\tau\eta}\left(\mathbb{E}\left[\mathcal{L}(\boldsymbol{x}^{t+1})\right] - \mathcal{L}(\boldsymbol{x}^t)\right) \\
&\quad\leq -\|\nabla \mathcal{L}(\boldsymbol{x}^t)\|^2 + 15\tau\eta L \sum_{i=1}^{m} w_i^2 \sigma^2 + \frac{\gamma}{2(1-\gamma)}\zeta^2 \\
&\quad+ \frac{\gamma\beta^2}{2(1-\gamma)}\mathbb{E}\|\nabla \mathcal{L}(\boldsymbol{x}^t)\|^2 + \frac{\eta^2\sigma^2L^2\left(\tau - 1\right)}{1-\gamma}.
\end{aligned} \tag{28}$$

Furthermore, if $\gamma \leq \frac{1}{2\beta^2+1}$, then we have $\frac{1}{1-\gamma} \leq 1 + \frac{1}{2\beta^2}$ and $\frac{\gamma\beta^2}{1-\gamma} \leq \frac{1}{2}$. Therefore the above result can be simplified as,

$$\begin{aligned}
&\frac{1}{\tau\eta}\left(\mathbb{E}\left[\mathcal{L}(\boldsymbol{x}^{t+1})\right] - \mathcal{L}(\boldsymbol{x}^t)\right) \leq -\frac{3}{4}\|\nabla \mathcal{L}(\boldsymbol{x}^t)\|^2 \\
&+ 15\tau\eta L \sum_{i=1}^{m} w_i^2 \sigma^2 + \eta^2\sigma^2L^2(\tau - 1)\left(1 + \frac{1}{2\beta^2}\right) \\
&+ 2\eta^2L^2\tau(\tau - 1)\left(1 + \frac{1}{2\beta^2}\right)\zeta^2
\end{aligned} \tag{29}$$

$$\begin{aligned}
&\leq -\frac{3}{4}\|\nabla \mathcal{L}(\boldsymbol{x}^t)\|^2 + 15\tau\eta L \sum_{i=1}^{m} w_i^2 \sigma^2 \\
&+ \frac{3}{2}\eta^2\sigma^2L^2(\tau - 1) + 3\eta^2L^2\tau(\tau - 1)\zeta^2.
\end{aligned} \tag{30}$$

Taking the average across all T communication rounds,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla \mathcal{L}(\boldsymbol{x}^t)\|^2 \leq \frac{4\left[\mathcal{L}(\boldsymbol{x}^0) - \mathcal{L}_{\inf}\right]}{3\eta\tau T}
\end{aligned} \tag{31}$$

$$+ 20\tau\eta L\sigma^2 \sum_{i=1}^{m} w_i^2 + 2\eta^2\sigma^2L^2(\tau - 1) + 4\eta^2L^2\tau(\tau - 1)\zeta^2.$$

For the ease of writing, we define $A = m\tau\sum_{i=1}^{m} w_i^2$, $B = \tau - 1$ and $C = \tau(\tau - 1)$, and then we derive

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla \mathcal{L}(\boldsymbol{x}^t)\|^2 \leq \frac{4\left[\mathcal{L}(\boldsymbol{x}^0) - \mathcal{L}_{\inf}\right]}{3\eta\tau T} \\
&+ \frac{20\eta L\sigma^2 A}{m} + 2\eta^2\sigma^2L^2B + 4\eta^2L^2C\zeta^2.
\end{aligned} \tag{32}$$

Since there is

$$\min_{t\in\mathbb{T}}\mathbb{E}\|\nabla \mathcal{L}(\boldsymbol{x}^t)\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla \mathcal{L}(\boldsymbol{x}^t)\|^2, \tag{33}$$

it holds that,

$$\begin{aligned}
&\min_{t\in\mathbb{T}}\mathbb{E}\|\nabla \mathcal{L}(\boldsymbol{x}^t)\|^2 \leq \frac{4\left[\mathcal{L}(\boldsymbol{x}^0) - \mathcal{L}_{\inf}\right]}{3\eta\tau T} \\
&+ \frac{20\eta L\sigma^2 A}{m} + 2\eta^2\sigma^2L^2B + 4\eta^2L^2C\zeta^2.
\end{aligned} \tag{34}$$

By setting $\eta = \sqrt{\frac{m}{\tau T}}$, we have

$$\begin{aligned}
&\min_{t\in\mathbb{T}}\mathbb{E}\|\nabla \mathcal{L}(\boldsymbol{x}^t)\|^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{m\tau T}}\right) + \mathcal{O}\left(\frac{A\sigma^2}{\sqrt{m\tau T}}\right) \\
&\quad + \mathcal{O}\left(\frac{mB\sigma^2}{\tau T}\right) + \mathcal{O}\left(\frac{mC\zeta^2}{\tau T}\right).
\end{aligned}$$

The proof of Theorem 1 is completed.

Comparison with vanilla FedAvg. Compared with the convergence analysis of FedAvg in [31], the above theorem theoretically indicates that SAFARI with unreliable communications can achieve the same asymptotic convergence rate as FedAvg with reliable communication network. Hence, the negative influence of communication unreliability is effectively controlled. In the next section, the experiment results that confirm our theoretical analysis are provided.

5 Experiments

We evaluate the proposed framework using different sparse algorithms with 50 clients. We train the ResNet-20 model [34] on the CIFAR-10 dataset, which contains 50,000 images for training and 10,000 images for testing. Specifically, the models are trained using Adam [35] optimizer with a learning rate of 0.001, batch size of 64 and tested using a batch size of 256. All of our experimental results are trained and evaluated using two NVIDIA-3090 GPUs with 24GB GPU RAM.

5.1 Performance of SAFARI on Non-IID Data Distribution

Evaluation Metrics. Since the each client has non-IID data to others, only simply computing the mean of local training accuracy and loss is unable to demonstrate the generalization of the global model. Therefore, we sample a small subset of each client's data and evaluate the testing accuracy and testing loss on the union of each subset. This gives us an indication of how well our global model is in a more comprehensive way, which also corresponds with the initialization stage of a typical FL setting [36].

To evaluate the generalization of our framework, we have compared the performance of SAFARI with three representative algorithms for neural network pruning. The sparsity level α is set to 60%, where 60% of model parameters will be pruned to 0. The selected pruning algorithms include: (1) MAG [13]: prunes the 60% smallest absolute values of the model parameters; (2) Synflow [13]: uses the synaptic saliency score to determine the importance of parameters in the network; (3) FedSpa [14]: gives evolutionary sparse masks to achieve personalized local models during FL training.

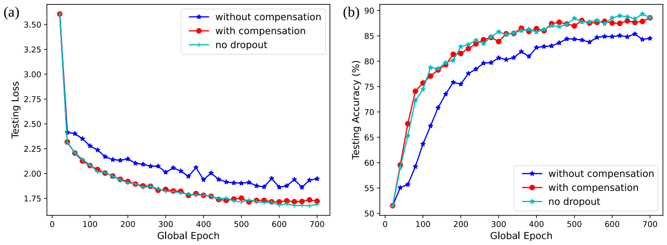

Fig. 2: Performance of SAFARI with FedAvg and MAG: (a) Testing accuracy; (b) Testing loss.

Following the balanced non-IID data partition setting [37] in FL, 50 total clients are divided into 2 groups equally, and each client contains 5 labels in CIFAR-10. Besides, local steps $\tau = 5$ and local learning rate $\eta = 0.001$ are set to perform the local sparse training in Algorithm 2. In addition, as addressed in Section 3, the successful transmission probability p_i^t is chosen as 0.6 for each link i in all communication rounds $t \in \{0, ..., T-1\}$, but each group has at least one client participating in each communication round. It is a simulation of real-world scenarios where the central server may fail to receive the updates from some clients due to unreliable communication. In our experiments, for the subset of absent gradients, SAFATRI will substitute them with alternatives, while other experiments without compensation will directly perform global aggregation without the absent gradients.

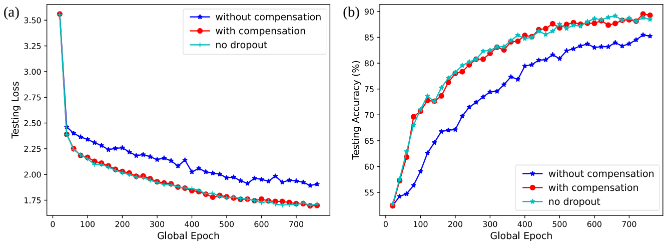

Fig. 3: Performance of SAFARI with FedAvg and Synflow: (a) Testing accuracy; (b) Testing loss.

Figure 2 shows the results of the proposed SAFARI framework with FedAvg as the global aggregation method and MAG as the local sparse training algorithm. Without the proposed similarity-based compensation scheme for bias reduction, the unreliable communication channel will degenerate the global model convergence. However, by introducing the compensation based on the similarity between sparse models, the convergence and performance of the global model are the same as under perfect communication.

Figure 3 evaluates SAFARI's performance with FedAvg and Synflow. It also compares the convergence behavior with respect to the number of iterations of global training with and without compensation, as well as the original experiments with no dropouts (i.e., every transmission succeeds). It is obvious that the training with compensation enabled by the Synflow sparse models has achieved nearly an identical rate of convergence and final accuracy as training without dropouts, which is far superior to the training without compensation. Similarly, experiments with FedAvg and FedSpa method also demonstrate the effectiveness of the proposed SAFARI framework, as shown in Figure 4.

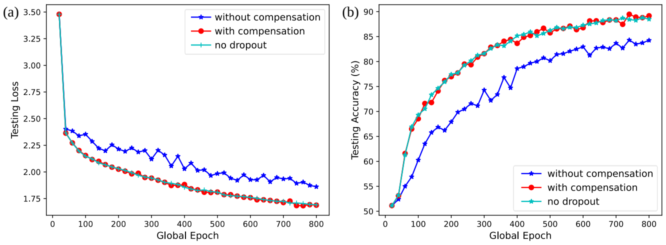

Fig. 4: Performance of SAFARI with FedAvg and FedSpa: (a) Testing accuracy; (b) Testing loss.

Moreover, we also investigate the performance of SAFARI with the FedProx [33] as the global aggregation algorithm. The results of Fedprox obtained with MAG, Synflow, and FedSpa as sparse methods are shown in Figure 9, Figure 10, and Figure 11 in appendix A respectively.

5.2 Validity of Similarity-based Compensation Scheme

In this section, the experiments are conducted to verify the validity of the proposed similarity-based compensation scheme. Following the lemmas in Section 4, the l_2-norm based distance of model parameters of two clients u, v is adopted in our experiment as the similarity function $s(\boldsymbol{x}_u, \boldsymbol{x}_v)$ in Algorithm 3:

$$s(\boldsymbol{x}_u, \boldsymbol{x}_v) := \|\boldsymbol{x}_u - \boldsymbol{x}_v\|. \tag{35}$$

Particularly, we display the final distance matrix ρ among all clients after the whole training is completed, as plotted in Figure 5. For this experiment, following the basic setting, clients 0 to 24 are in Group 1 and have the same label split, while clients 25 to 49 are in Group 2. The darker-colored areas in the upper left and lower right corners indicate that the distance between sparse models computed by clients in the same group is relatively small. It is aligned with the fact that the underlying data distributions among clients in the same group are of higher similarity. By contrast, the areas in the lower left and upper right corners of this figure correspond to the model distances among clients from different groups, and the light colors indicate large distance values and low similarity. This similarity matrix proves that the proposed SAFARI tends to compensate for missing model updates with other clients' updates from the same group.

5.3 Evaluation of Stability

In this section, we evaluate the validity and stability of SAFARI by changing two decisive hyperparameters: the

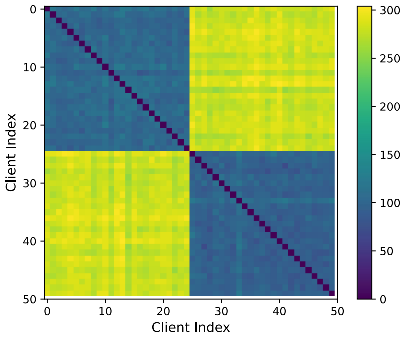

Fig. 5: The final distance matrix.

successful transmission probability P and the sparsity level α. Specifically, only these two hyperparameters are changed and the other experiment settings follow the MAG test with FedAvg.

5.3.1 Successful Transmission Probability

For simplicity, we denote the situation where the transmission success probability for all the communication links is equal to a as $P = a$. To explore the validity of SAFARI under different P, five different P values are selected and the results are shown in Figure 6 and Figure 7. In Figure 6, FL training without compensation largely depends on reliable communication conditions. When the communication is unreliable, e.g. $P < 1.0$, the convergence and model performance are obviously impacted. By contrast, as the successful transmission probability P changes, the convergence and model accuracy of FL training with our proposed compensation remains the same, as shown in Figure 7. The results prove the SAFARI framework is robust to varying and dynamic communication reliability, and is capable of alleviating the impact of unreliable communication.

5.3.2 Sparsity Level

Figure 8 shows the test loss and testing accuracy of SAFARI with different sparsity levels from 0 to 0.8, where $\alpha = 0$ refers to no sparsification. When the sparsity level α varies, the impact on the testing accuracy and testing loss is limited, especially when the sparsity level is lower than 0.7. Therefore, SAFARI ensures satisfactory convergence and model performance with a moderate sparsity level, which is sufficient to achieve significant transmission reduction at the same time.

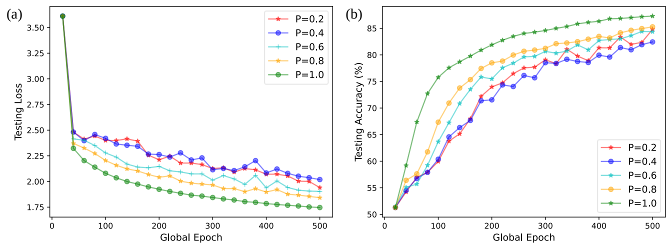

Fig. 6: Performance under different transmission probability settings (without compensation): (a) Testing accuracy; (b) Testing loss.

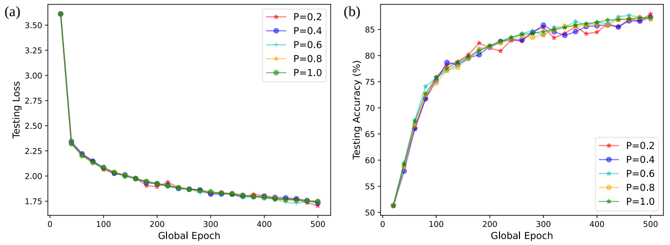

Fig. 7: Performance under different transmission probability settings (with compensation): (a) Testing accuracy; (b) Testing loss.

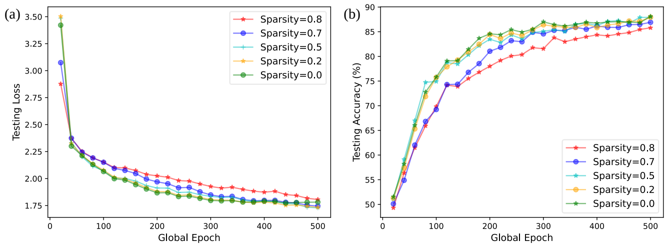

Fig. 8: Performance under different sparsity settings: (a) Testing accuracy; (b) Testing loss.

6 Conclusion

In this paper, we propose a sparsity-enabled robust FL framework, named as SAFARI, which can reduce communication overhead by local sparse learning, and meanwhile rectify the aggregation bias resulting from unreliable communications with unknown and potentially time-varying unreliability characteristics. Our theoretical analysis with respect to sparse models demonstrates that the similarity properties of client models are preserved under sparsity, and thus the proposed SAFARI algorithm with the similarity-based compensation can achieve the same asymptotic convergence rate as FedAvg with reliable communications. The experiments with CIFAR10 dataset and several representative sparse algorithms show that SAFARI can not only save up to 60% communication overhead but also consistently outperforms baselines by achieving fast and stable convergence under unreliable communications. Future work includes extending our work to consider more complex factors, such as fading and shadowing channels, upstream compression, extremely heterogeneous data, etc. We believe it's possible to further utilize the similarity of sparse models to correct for corrupted messages caused by fading channels. Moreover, we reckon some state-of-the-art methods specifically designed for the heterogeneous issue, such as knowledge distillation, can also be applied together with our framework to alleviate the underlying data heterogeneity, and thus are promising for more general applications.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant No.2022ZD0160504, by Tsinghua Shenzhen International Graduate School-Shenzhen Pengrui Young Faculty Program of Shenzhen Pengrui Foundation (No. SZPR2023005), and by Tsinghua-Toyota Joint

Research Institute inter-disciplinary Program and Tsinghua University (AIR)-Asiainfo Technologies (China) Inc. Joint Research Center under grant No. 20203910074. We would also like to thank anonymous reviewers for their insightful comments.

References

[1] Z. Qin, G. Y. Li, and H. Ye, "Federated learning and wireless communications," *IEEE Wireless Communications*, vol. 28, no. 5, pp. 134–140, 2021.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, USA, April 2017, pp. 1273–1282.

[3] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, pp. 1–19, 2019.

[4] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *ACM SIGSAC Conference on Computer and Communications Security*, Vienna, Austria, May 2016, p. 308–318.

[5] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *ACM SIGSAC Conference on Computer and Communications Security*, Dallas, Texas, USA, May 2017, p. 1175–1191.

[6] Y. Liu, Y. Kang, C. Xing, T. Chen, and Q. Yang, "A secure federated transfer learning framework," *IEEE Intelligent Systems*, vol. 35, no. 4, pp. 70–82, 2020.

[7] X. Ma, M. Qin, F. Sun, Z. Hou, K. Yuan, Y. Xu, Y. Wang, Y.-K. Chen, R. Jin, and Y. Xie, "Effective model sparsification by scheduled grow-and-prune methods," in *International Conference on Learning Representations*, April 2022.

[8] B. Woodworth, K. K. Patel, S. Stich, Z. Dai, B. Bullins, B. Mcmahan, O. Shamir, and N. Srebro, "Is local SGD better than minibatch SGD?" in *International Conference on Machine Learning*, July 2020, pp. 10 334–10 343.

[9] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich, "A unified theory of decentralized SGD with changing topology and local updates," in *International Conference on Machine Learning*, July 2020, pp. 5381–5393.

[10] A. Khaled, K. Mishchenko, and P. Richtárik, "Tighter theory for local SGD on identical and heterogeneous data," in *International Conference on Artificial Intelligence and Statistics*, 2020, pp. 4519–4529.

[11] H. Ye, L. Liang, and G. Y. Li, "Decentralized federated learning with unreliable communications," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–13, 2022.

[12] Y. Fraboni, R. Vidal, L. Kameni, and M. Lorenzi, "Clustered sampling: Low-variance and improved representativity for clients selection in federated learning," in *International Conference on Machine Learning*, July 2021, pp. 3407–3416.

[13] H. Tanaka, D. Kunin, D. L. Yamins, and S. Ganguli, "Pruning neural networks without any data by iteratively conserving synaptic flow," in *Advances in Neural Information Processing Systems*, vol. 33, December 2020, pp. 6377–6389.

[14] T. Huang, S. Liu, L. Shen, F. He, W. Lin, and D. Tao, "Achieving personalized federated learning with sparse local models," *arXiv preprint arXiv:2201.11380*, 2022.

[15] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Advances in Neural Information Processing Systems*, vol. 30, CA, USA, December 2017, pp. 1709–1720.

[16] D. Jhunjhunwala, A. Gadhikar, G. Joshi, and Y. C. Eldar, "Adaptive quantization of model updates for communication-efficient federated learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, June 2021, pp. 3110–3114.

[17] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGD: Compressed optimisation for non-convex problems," in *International Conference on Machine Learning*, Stockholm, Sweden, July 2018, pp. 560–569.

[18] N. Strom, "Scalable distributed dnn training using commodity gpu cloud computing," in *Annual Conference of the International Speech Communication Association*, Dresden, Germany, September 2015.

[19] J. Wangni, J. Wang, J. Liu, and T. Zhang, "Gradient sparsification for communication-efficient distributed optimization," in *Advances in Neural Information Processing Systems*, vol. 31, Montreal, CANADA, December 2018, pp. 1299–1309.

[20] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, "Sparsified SGD with memory," in *Advances in Neural Information Processing Systems*, vol. 31, Montreal, CANADA, December 2018, pp. 1–12.

[21] S. P. Karimireddy, Q. Rebjock, S. Stich, and M. Jaggi, "Error feedback fixes signsgd and other gradient compression schemes," in *International Conference on Machine Learning*, California, USA, June 2019, pp. 3252–3261.

[22] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-iid data," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 9, pp. 3400–3413, 2019.

[23] D. C. Mocanu, E. Mocanu, P. Stone, P. H. Nguyen, M. Gibescu, and A. Liotta, "Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science," *Nature Communications*, vol. 9, no. 1, pp. 1–12, 2018.

[24] N. Lee, T. Ajanthan, and P. Torr, "Snip: Single-shot network pruning based on connection sensitivity," in *International Conference on Learning Representations*, New Orleans, USA, May 2019.

[25] T. Dettmers and L. Zettlemoyer, "Sparse networks from scratch: Faster training without losing performance," *arXiv preprint arXiv:1907.04840*, 2019.

[26] H. You, C. Li, P. Xu, Y. Fu, Y. Wang, X. Chen, R. G. Baraniuk, Z. Wang, and Y. Lin, "Drawing early-bird tickets: Towards more efficient training of deep networks," *arXiv preprint arXiv:1909.11957*, 2019.

[27] U. Evci, T. Gale, J. Menick, P. S. Castro, and E. Elsen, "Rigging the lottery: Making all tickets winners," in *International Conference on Machine Learning*, July 2020, pp. 2943–2952.

[28] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning*, July 2020, pp. 5132–5143.

[29] S. J. Reddi, J. Konečnỳ, P. Richtárik, B. Póczós, and A. Smola, "Aide: Fast and communication efficient distributed optimization," *arXiv preprint arXiv:1608.06879*, 2016.

[30] T. Murata and T. Suzuki, "Bias-variance reduced local SGD for less heterogeneous federated learning," *arXiv preprint arXiv:2102.03198*, 2021.

[31] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in *Advances in Neural Information Processing Systems*, vol. 33, Vancouver, Canada, Dec. 2020, pp. 7611–7623.

[32] H. You, C. Li, P. Xu, Y. Fu, Y. Wang, X. Chen, Z. Wang, R. G. Baraniuk, and Y. Lin, "Drawing early-bird tickets: Towards more efficient training of deep networks," in *International Conference on Learning Representations*, Addis Ababa, Ethiopia, April 2020.

[33] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine learning and systems*, vol. 2, pp. 429–450, 2020.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA, June 2016, pp. 770–778.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, San Diego, CA, USA, May 2015.

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

[37] E. Diao, J. Ding, and V. Tarokh, "Heterofl: Computation and communication efficient federated learning for heterogeneous clients," in *International Conference on Learning Representations*, 2020.

Yuzhu Mao received the B.E. degree in computer science from Wuhan University, Wuhan, China, in 2020. Yuzhu Mao is currently pursuing her M.S. degree in Data Science and Information Technology at Smart Sensing and Robotics (SSR) group, Tsinghua University. Her research interests include Federated Learning, Internet of Things (IoTs), and Multi-agent Systems.

Zihao Zhao received his B.S. degree in University of Electronic Science and Technology of China (UESTC) in 2021. He is currently pursuing his M.S. degree in Data Science and Information Technology at Smart Sensing and Robotics (SSR) group, Tsinghua University. His research interests include Internet of Things (IoTs), Federated Learning, and Machine Learning.

Meilin Yang received the BSEE degree from the Wuhan University, China in 2021. Meilin Yang is currently a master student at the Tsinghua-Berkeley Shenzhen Institude (TBSI). Her interests include Federated Learning, Deep Hashing, Graph Neural Network and Knowledge Graphs.

Le Liang (S'13-M'19) received the B.E. degree in information engineering from Southeast University, Nanjing, China, in 2012, the M.A.Sc degree in electrical engineering from the University of Victoria, Victoria, BC, Canada, in 2015, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, in 2018. From 2019 to 2021, he was a Research Scientist at Intel Labs, Hillsboro, OR. Since 2021, he has been with the School of Information Science and Engineering, Southeast University, Nanjing, China, where he is currently a Professor with the National Mobile Communications Research Laboratory. His main research interests are in wireless communications, signal processing, and machine learning.

Dr. Liang serves as an Editor for the IEEE Communications Letters and as an Associate Editor for the IEEE JSAC Series on Machine Learning in Communications and Networks. He is a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society. He received the Best Paper Award of IEEE/CIC ICCC in 2014 and was named an Exemplary Reviewer of the IEEE Wireless Communications Letters in 2018.

Yang Liu is an associate professor with the Institute for AI Industry Research, Tsinghua University. Before joining Tsinghua, she was the principal researcher and research team lead at WeBank. Her research interests include federated learning, machine learning, multi-agent systems, statistical mechanics and AI industrial applications. Her research work was recognized with multiple awards, such as AAAI Innovation Award and CCF Technology Award. She is also named as Innovators on Privacy-Preserving Computation by MIT Technology Review China.

Wenbo Ding received the BS and PhD degrees (Hons.) from Tsinghua University in 2011 and 2016, respectively. He worked as a postdoctoral research fellow at Georgia Tech under the supervision of Professor Z. L. Wang from 2016 to 2019. He is now an associate professor and PhD supervisor at Tsinghua-Berkeley Shenzhen Institute, Institute of Data and Information, Shenzhen International Graduate School, Tsinghua University, where he leads the Smart Sensing and Robotics (SSR) group. His research interests are diverse and interdisciplinary, which include self-powered sensors, energy harvesting, and wearable devices for health and robotics with the help of signal processing, machine learning, and mobile computing. He has received many prestigious awards, including the Gold Medal of the 47th International Exhibition of Inventions Geneva and the IEEE Scott Helt Memorial.

Tian Lan received the B.A.Sc. degree from the Tsinghua University, China in 2003, the M.A.Sc. degree from the University of Toronto, Canada, in 2005, and the Ph.D. degree from the Princeton University in 2010. Dr. Lan is currently a full Professor of Electrical and Computer Engineering at the George Washington University. His research interests include network optimization, algorithms, and machine learning. He received the Meta Research Award in 2021, SecureComm Best Paper Award in 2019, SEAS Faculty Recognition Award in 2018, Hegarty Faculty Innovation Award in 2017, AT&T VURI Award in 2015, IEEE INFOCOM Best Paper Award in 2012, Wu Prizes for Excellence at Princeton University in 2010, IEEE GLOBECOM Best Paper Award in 2009, and IEEE Signal Processing Society Best Paper Award in 2008.

Xiao-Ping Zhang received B.S. and Ph.D. degrees from Tsinghua University, in 1992 and 1996, respectively, both in Electronic Engineering. He holds an MBA in Finance, Economics and Entrepreneurship with Honors from the University of Chicago Booth School of Business, Chicago, IL.

He is a Professor with Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, and with the Department of Electrical, Computer and Biomedical Engineering, Toronto Metropolitan University (Formerly Ryerson University), Toronto, ON, Canada, where he is the Director of the Communication and Signal Processing Applications Laboratory. He is cross-appointed to the Finance Department at the Ted Rogers School of Management, Toronto Metropolitan University. He was a Visiting Scientist with the Research Laboratory of Electronics, Massachusetts Institute of Technology. His research interests include sensor networks and IoT, image and multimedia content analysis, machine learning, statistical signal processing, and applications in big data, finance, and marketing.

Dr. Zhang is Fellow of the Canadian Academy of Engineering, Fellow of the Engineering Institute of Canada, Fellow of the IEEE, a registered Professional Engineer in Ontario, Canada, and a member of Beta Gamma Sigma Honor Society. He is the general Co-Chair for the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2021. He is the general co-chair for 2017 and 2019 GlobalSIP Symposium on Signal, Information Processing and AI for Finance and Business. He was an elected Member of the ICME steering committee. He is the General Chair for the IEEE International Workshop on Multimedia Signal Processing, 2015. He is Editor-in-Chief for the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING. He is Senior Area Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING. He served as Senior Area Editor the IEEE TRANSACTIONS ON SIGNAL PROCESSING and Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and the IEEE SIGNAL PROCESSING LETTERS. He is the Chair of the IEEE Signal Processing Society Technical Committee on Image, Video, and Multidimensional Signal Processing (IVMSP). He received Sarwan Sahota Ryerson Distinguished Scholar Award – the Ryerson University highest honor for scholarly, research and creative achievements. He is an IEEE Distinguished Lecturer of the IEEE Signal Processing Society, and of the IEEE Circuits and Systems Society.