Byzantine-resilient Federated Learning With Adaptivity to Data Heterogeneity

Shiyuan Zuo, Xingrun Yan, Rongfei Fan, Han Hu, Hangguan Shan, and Tony Q. S. Quek

Abstract

This paper deals with federated learning (FL) in the presence of malicious Byzantine attacks and data heterogeneity. A novel Robust Average Gradient Algorithm (RAGA) is proposed, which leverages the geometric median for aggregation and can freely select the round number for local updating. Different from most existing resilient approaches, which perform convergence analysis based on strongly-convex loss function or homogeneously distributed dataset, we conduct convergence analysis for not only strongly-convex but also non-convex loss function over heterogeneous dataset. According to our theoretical analysis, as long as the fraction of dataset from malicious users is less than half, RAGA can achieve convergence at rate $\mathcal{O}(1/T^{2/3-\delta})$ where T is the iteration number and $\delta \in (0, 2/3)$ for non-convex loss function, and at linear rate for strongly-convex loss function. Moreover, stationary point or global optimal solution is proved to obtainable as data heterogeneity vanishes. Experimental results corroborate the robustness of RAGA to Byzantine attacks and verifies the advantage of RAGA over baselines on convergence performance under various intensity of Byzantine attacks, for heterogeneous dataset.

Index Terms

Federated learning, Byzantine attack, data heterogeneity, robust aggregation, convergence analysis.

S. Zuo, X. Yan, and R. Fan are with the School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing 100081, China (e-mail: {zuoshiyuan, 3220221473, fanrongfei}@bit.edu.cn); H. Hu is with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: hhu@bit.edu.cn); H. Shan is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China, and also with Zhejiang Provincial Key Laboratory of Information Processing, Communication and Networking, Hangzhou, China (e- mail: hshan@zju.edu.cn); T. Q. S. Quek is with the Singapore University of Technology and Design, Singapore 487372, and also with the Yonsei Frontier lab, Yonsei University, South Korea (e-mail: tonyquek@sutd.edu.sg).

I. Introduction

The rapid growth of intelligent applications in recent decade promotes a broad adoption of them on various devices or equipments, such as mobile phones, wearable devices, and autonomous vehicles, etc [1]–[3]. To enable one intelligent application, raw data from various subscribed users has to be trained together to generate a uniform model. From the perspective of subscribed users, who may distribute broadly, it has a risk of privacy leakage if they offload their raw data to a central server for training [4], [5]. Federated learning (FL) is a distributed training framework and has emerged as a promising solution for this dilemma [6], [7]. In FL, model training is completed by exchanging something about model parameters, such as the model parameter itself or the gradient of loss function with model parameters, between each involving user and one central server iteratively. In each round of iteration, the central server aggregates the model parameters from every user and subsequently broadcasts the aggregated one to all the users. After receiving the broadcasted model parameter, each user performs the role of local updating based on its own data. In such a procedure, there is no exchange of raw data between any user and the central server [8], [9].

However, distributed training like FL has to face robust issue because of the involvement of multiple users. To be specific, due to data corruption, device malfunctioning, or malicious attacks at some users, the information about the model parameter to be uploaded to the central server by these abnormal users may deviate from the expected one [10], [11]. In such a case, the group of abnormal users are called as *Byzantine users*, and the action of Byzantine users is referred to as *Byzantine attacks* [12], [13]. As a comparison, the group of normal users are called as *Honest users*. For a Byzantine attack, it is general to assume the uncertainty of the identity and the population of Byzantine users. What is more, the attack initiated by Byzantine users could be arbitrarily malicious [14], [15]. With such a setup, the training performance of FL will be surely degraded and aggregation strategies resistant to Byzantine attacks have been concerned in literature [16].

On the other hand, the training process of FL also faces the challenges of data heterogeneity among multiple users. This heterogeneity comes from the diverse distribution of local data at multiple users, since the local data may be collected by each FL user under its specific environment, habit, or preference [17]. In such a case, data distributions over multiple users are always assumed to be non independently and identically distributed (non-IID). The presence of data heterogeneity has already been demonstrated to cause local drift of user model, and thus, degradation of the global model's performance. How to be adaptive to data heterogeneity is also an important issue and has been studied extensively in the area of FL [18].

A. Related Works

In related works, aggregation strategies are designed to be robust to the Byzantine attack under various setup of system assumptions, whose performance is then justified by convergence analysis and/or experimental results. The system assumptions usually relate to the nature of the data (IID or non-IID), the nature of the loss function (strongly-convex or non-convex), and the way of local updating (one-step or multiple steps), etc. For these assumptions, the data heterogeneity and non-convex loss function make FL framework more general, and multiple steps of local updating waives each user from frequent communication with the central server. The Byzantine-resilient aggregation strategies in the literature of FL under diverse system assumptions have been summarized in Table I.

As of the aggregation strategies in related literature, by leveraging median [16], trimmed mean [19], or iterative filtering [20], the associated FL system can tolerate a small fraction of Byzantine users. To be more robust to Byzantine attacks, some methods such as Krum [21], Federated Learning Over the Air (FLOA) [22], Robust Stochastic Aggregation (RSA) [23], and Robust Aggregating Normalized Gradient Method (RANGE) [24], have been proposed. Specifically, Krum selects a stochastic gradient as the global one, which has the shortest cumulative distance from a group of stochastic gradients that are most closely distributed, FLOA requires each user to normalize the associated local gradient to be a vector with zero mean and unit variance before offloading it to the central server, RSA penalizes the difference between local model parameters and global model parameters to isolate Byzantine users, and RANGE takes the coordinate-wise median of uploaded stochastic gradients as the aggregated one.

Although the above aggregation strategies can be resistant to Byzantine attacks at some extent, they may still fail to work in front of heavy Byzantine attacks: The attacks with the fraction of Byzantine users being large or the amplitude of uploaded vectors by Byzantine users being high. Recently, geometric median appears as a good candidate for aggregation because of its ability to naturally tolerate a larger fraction of strong Byzantine attacks [28]. Based on geometric median, some algorithms such as Robust Federated Aggregation (RFA) [25], Byzantine attack resilient distributed Stochastic Average Gradient Algorithm (Byrd-SAGA) [26], and Byzantine-

TABLE I: List of references on the convergence of FL under Byzantine attacks.

References	Algorithm	Function Type	Data Heterogeneity	Local Updating	Optimality Gap	Other Assumption
[16]	median	-	IID	one step	-	-
[19]	trimmed mean	non-convex strongly-convex	IID	one step	non-zero	Bounded gradient
[20]	iterative filtering	strongly-convex	IID	one step	non-zero	-
[21]	Krum	non-convex	IID	one step	non-zero	-
[22]	FLOA	non-convex	IID	one step	zero	Bounded gradient Unbiased gradient
[23]	RSA	strongly-convex	non-IID	one step	non-zero	Bounded gradient
[24]	RANGE	non-convex strongly-convex	IID	one step	non-zero	-
[25]	RFA	strongly-convex	non-IID	multiple steps	non-zero	-
[26]	Byrd-SAGA	strongly-convex	IID	one step	non-zero	Bounded gradient Unbiased gradient
[27]	BROADCAST	strongly-convex	IID	one step	non-zero	Bounded gradient Unbiased gradient
This work	RAGA	non-convex strongly-convex	non-IID	multiple steps	non-zero	Bounded gradient

RObust Aggregation with gradient Difference Compression And STochastic variance reduction (BROADCAST) [27], have been proposed. To be specific, geometric median is leveraged to aggregate the uploaded vectors for RFA, Byrd-SAGA, and BROADCAST. Differently, each user in RFA selects the trimmed mean of the model parameters over multiple local updating as the uploaded vector, while Byrd-SAGA and BROADCAST utilize the SAGA [29] method to generate the vectors for aggregation first and then upload them in a lossless or compressive way.

In aforementioned Byzantine-resilient aggregation strategies, there is little work claimed to be adaptive to data heterogeneity, which however is a non-negligible issue. RFA in [25] is an exception. However, RFA only investigates strongly-convex loss function. With such a convexity setup, RFA shows that it is can converge to a neighborhood of the optimal solution only if the number of local updates for each involving user is larger than then a threshold so as to fulfill a necessary inequality. However, this restricts the flexibility of selecting the number of local updates and maybe computationally intensive when these users have limited computation capability.

B. Motivations and Contributions

In this paper we investigate Byzantine-resilient FL with adaptivity to data heterogeneity, which was rarely studied in literature. Different from RFA, which is the unique work dealing with both Byzantine attack and data heterogeneity in FL by now but only considers strongly-convex loss function, we investigate not only strongly-convex but also non-convex loss function. Moreover, with regard to strongly-convex loss function, not like RFA that requires the step number of local updates to be above some threshold (which is a prerequisite for promising convergence in RFA), we assume arbitrary selection of it, which is more general. With our featured system setup, we propose a novel FL structure, RAGA. Specifically, the RAGA leverages the geometric median to aggregate the local gradients offloaded by the users after multiple rounds of local updating. Associated analysis of convergence and robustness are then followed for both non-convex and strongly-convex loss functions. Our main contributions are summarized as follows:

- Algorithmically, we propose a new FL aggregation strategy, named RAGA, which has the flexibility of arbitrary selection of local update's step number, and is robust to both Byzantine attack and data heterogeneity.
- Theoretically, we established a convergence guarantee for our proposed RAGA under not only non-convex but also strongly-convex loss function. Through rigorous proof, we show that the established convergence can be promised as long as the fraction of dataset from Byzantine users is less than half. The achievable convergence rate can be at $\mathcal{O}(1/T^{2/3-\delta})$ where T is the iteration number and $\delta \in (0, 2/3)$ for non-convex loss function, and at linear for strongly-convex loss function. Moreover, stationary point and global optimal solution are shown to be obtainable as data heterogeneity vanishes, for non-convex and strongly-convex loss function, respectively.
- Empirically, we conduct extensive experiments to evaluate the performance of our proposed RAGA. With non-IID dataset, our proposed RAGA is verified to be robust to various Byzantine attack level. Moreover, compared with baseline methods, our proposed RAGA can achieve higher test accuracy and lower training loss.

C. Organization and Notation

The rest of this paper is organized as follows: Section II introduces the system model and formulates the problem to be solved in this paper, Section III presents our proposed RAGA. In Section IV, the analysis of convergence and robustness for our proposed RAGA under non-convex and strongly-convex loss function are disclosed. Experimental results are illustrated in Section V, followed by conclusion remarks in Section VI.

Notation: We use plain and bold letters to represent scalars and vectors respectively, e.g., a and b are scalars, $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors. $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ denotes the inner product of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, $\nabla F(\boldsymbol{x}_0)$ represents the gradient vector of function $F(\boldsymbol{x})$ at $\boldsymbol{x} = \boldsymbol{x}_0$, $\|\boldsymbol{x}\|$ stands for the Euclidean norm $\|\boldsymbol{x}\|_2$, $\mathbb{E}\{a\}$ and $\mathbb{E}\{\boldsymbol{x}\}$ implies the expectation of random variable a and random vector $\boldsymbol{x}$ respectively, and $a \propto b$ means that a is proportional to b.

II. Problem Statement

Consider a FL system with one central server and M users, which composite the set $\mathcal{M} \triangleq \{1, 2, ..., M\}$. At the side of mth user, it has a dataset $\mathcal{S}_m$, which has S_m ground-true labels. In the FL system, the learning task is to train a model parameter $\boldsymbol{w} \in \mathbb{R}^p$ to minimize the global loss function, denoted as $F(\boldsymbol{w})$, for approximating the data labels of all the users. Specifically, we need to solve the following problem

Problem 1:

$$\min_{\boldsymbol{w} \in \mathbb{R}^p} F(\boldsymbol{w})$$

In Problem 1, the global loss function $F(\boldsymbol{w})$ is defined as

$$F(\boldsymbol{w}) \triangleq \frac{1}{\sum_{m \in \mathcal{M}} S_m} \sum_{m \in \mathcal{M}} \sum_{\boldsymbol{s} \in \mathcal{S}_m} f(\boldsymbol{w}, \boldsymbol{s}) \tag{1}$$

where $f(\boldsymbol{w}, \boldsymbol{s})$ denotes the loss function for data sample $\boldsymbol{s}$. For convenience, we define the local loss function of mth user as

$$F_m(\boldsymbol{w}) \triangleq \frac{1}{S_m} \sum_{\boldsymbol{s} \in \mathcal{S}_m} f(\boldsymbol{w}, \boldsymbol{s}), \tag{2}$$

then the global loss function $F(\boldsymbol{w})$ can be rewritten as

$$F(\boldsymbol{w}) = \sum_{m \in \mathcal{M}} \frac{S_m}{\sum_{i \in \mathcal{M}} S_i} F_m(\boldsymbol{w}). \tag{3}$$

For the global loss function $F(\boldsymbol{w})$ and local loss function $F_m(\boldsymbol{w})$ for $m \in \mathcal{M}$, we adopt assumptions from the following ones.

Assumption 1 (Lipschitz Continuity): The loss function $f(\boldsymbol{w}, \boldsymbol{s})$ has L-Lipschitz continuity [30], [31], i.e., for $\forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^p$, there is

$$f(\boldsymbol{w}_1, \boldsymbol{s}) - f(\boldsymbol{w}_2, \boldsymbol{s}) \leqslant \langle \nabla f(\boldsymbol{w}_2, \boldsymbol{s}), \boldsymbol{w}_1 - \boldsymbol{w}_2 \rangle + \frac{L}{2} \left\| \boldsymbol{w}_1 - \boldsymbol{w}_2 \right\|^2, \tag{4}$$

which is equivalent to the following inequality,

$$\left\| \nabla f(\boldsymbol{w}_1, \boldsymbol{s}) - \nabla f(\boldsymbol{w}_2, \boldsymbol{s}) \right\| \leqslant L \left\| \boldsymbol{w}_1 - \boldsymbol{w}_2 \right\|. \tag{5}$$

With (4) or (5), it can be derived that the global loss function $F(\boldsymbol{w})$ and local loss function $F_m(\boldsymbol{w}), \forall m \in \mathcal{M}$ both have L-Lipschitz continuity.

Assumption 2 (Strongly-Convex): The loss function $f(\boldsymbol{w}, \boldsymbol{s})$ is μ strongly-convex [30], i.e., for $\forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^p$, there is

$$\left\| \nabla f(\boldsymbol{w}_1, \boldsymbol{s}) - \nabla f(\boldsymbol{w}_2, \boldsymbol{s}) \right\| \geqslant \mu \left\| \boldsymbol{w}_1 - \boldsymbol{w}_2 \right\|. \tag{6}$$

Similarly, with (6), $F(\boldsymbol{w})$ and $F_m(\boldsymbol{w}), \forall m \in \mathcal{M}$ can be derived to be both μ strongly-convex.

Assumption 3 (Local Unbiased Gradient): For $\forall m \in \mathcal{M}$, suppose a subset of dataset $\mathcal{S}_m$ is selected randomly, which is denote as ξ_m, define

$$F_m(\boldsymbol{w}; \xi_m) \triangleq \frac{1}{|\xi_m|} \sum_{\boldsymbol{s} \in \xi_m} f(\boldsymbol{w}, \boldsymbol{s}), \tag{7}$$

the unbiased gradient [30], [31] implies that

$$\mathbb{E}\left\{ \nabla F_m(\boldsymbol{w}; \xi_m) \right\} = \nabla F_m(\boldsymbol{w}), \forall m \in \mathcal{M}. \tag{8}$$

Assumption 4 (Bounded Inner Variance): For every user m and $\boldsymbol{w} \in \mathbb{R}^p$, the variance of stochastic local gradient $\nabla F_m(\boldsymbol{w}; \xi_m)$ is upper-bounded [30], [31] by

$$\mathrm{Var}(\nabla F_m(\boldsymbol{w}; \xi_m)) \leqslant \sigma^2, \forall m \in \mathcal{M}. \tag{9}$$

Assumption 5 (Bounded Gradient): For every user m and $\boldsymbol{w} \in \mathbb{R}^p$, the ideal local gradient $\nabla F_m(\boldsymbol{w})$ is upper-bounded [30], [31] by

$$\left\| \nabla F_m(\boldsymbol{w}) \right\| \leqslant G, \forall m \in \mathcal{M}. \tag{10}$$

In traditional FL, iterative interactions between a group of M users and a central server are performed to update the gradient parameter $\nabla F(\boldsymbol{w})$ until convergence. For tth round of iteration, a conventional operation procedure can be given as the following steps:

Step 1 (Local Update): For mth user, it calculates the local gradient parameter $\nabla F_m(\boldsymbol{w}^t)$ from its local dateset $\mathcal{S}_m$ and global model parameter $\boldsymbol{w}^t$.

Step 2 (Aggregation and Broadcasting): Each user sends its local gradient parameter $\nabla F_m(\boldsymbol{w}^t)$ to the central server. The central server then aggregates all the local gradient vectors into a common one, denoted as $\boldsymbol{w}^{t+1}$, as follows,

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta \cdot \sum_{m=1}^{M} \nabla F_m(\boldsymbol{w}^t), \tag{11}$$

then the central server broadcasts $\boldsymbol{w}^{t+1}$ to all the users.

With some combination of previously claimed assumptions, the above iterative operation may help to reach the optimal or convergent solution of Problem 1 if each user sends trustworthy message to the central server. However, under Byzantine attacks, the main challenge of solving Problem 1 comes from the fact that the Byzantine attackers can collude and send arbitrary malicious messages to the central server so as to bias the optimization process. We aspire to develop a robust FL algorithm to address this issue in next section.

III. Proposed Aggregation Strategy: Robustness Average Gradient Algorithm (RAGA)

Before the introduction of the RAGA, we first explain the scenario of Byzantine attack. Assume there are B Byzantine users out of M users, which compose the set of $\mathcal{B}$. Any Byzantine user can send an arbitrary vector $\star, \star \in \mathbb{R}^p$ to the central server. Suppose $\boldsymbol{z}_m^t$ is the real vector uploaded by mth user to the central server in tth round of iteration, then there is

$$\boldsymbol{z}_m^t = \begin{cases} \boldsymbol{g}_m^t, & m \in \mathcal{M} \setminus \mathcal{B} \\ \star, & m \in \mathcal{B} \end{cases} \tag{12}$$

where g_m^t denotes the vector to be uploaded by honest user ($m \in \mathcal{M} \setminus \mathcal{B}$). Additionally, we also assume $\sum_{m \in \mathcal{B}} S_m < \frac{1}{2} \sum_{m \in \mathcal{M}} S_m$. In related works [26], by implicitly supposing S_m to be identical for every $m \in \mathcal{M}$, it is usually assumed $B < \frac{1}{2} M$, which is a special realization of our assumption.

To reduce communication burden due to frequent interactions between M users and the central server and overcome the Byzantine attacks, we propose the RAGA, which runs multi-local updates and aggregates all the uploaded vectors in a robust way. Specifically, in tth round of iteration:

Step 1 (Multi-Local Update): For any honest user m, $m \in \mathcal{M} \setminus \mathcal{B}$. Denote $\boldsymbol{w}_m^{t,k}$ as the local model parameter of mth user for kth update and $\eta_m^{t,k}$ as the associated learning rate. By implementing stochastic gradient descend (SGD) method, there is

$$\boldsymbol{w}_m^{t,k} = \boldsymbol{w}_m^{t,k-1} - \eta_m^{t,k} \cdot \nabla F_m(\boldsymbol{w}_m^{t,k-1}; \xi_m^{t,k}), k = 1, 2, \ldots, K^t \tag{13}$$

where $\xi_m^{t,k}$ is a randomly selected subset of dataset $\mathcal{S}_m$ and $\boldsymbol{w}_m^{t,0} = \boldsymbol{w}^t$. The $\boldsymbol{w}^t$ is received from $t-1$ round of iteration. Specially, for the sake of robustness and convergence, we set $\boldsymbol{z}_m^t = \boldsymbol{g}_m^t = \frac{1}{K^t} \sum_{k=1}^{K^t} \nabla F_m(\boldsymbol{w}_m^{t,k-1}; \xi_m^{t,k})$ for $m \in \mathcal{M} \setminus \mathcal{B}$.

Step 2 (Aggregation and Broadcasting): For the mth user, it uploads its own vector $\boldsymbol{z}_m^t$ to the central server for $m \in \mathcal{M}$. To ensure the robustness of the FL system, we use the numerical geometric median of all the uploaded vectors, which is defined as

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta^t \cdot \boldsymbol{z}^t \tag{14}$$

where η^t represents the global learning rate for tth round of iteration and $\boldsymbol{z}^t$ represents the numerical estimation of the geometric median of $\{\boldsymbol{z}_m^t | m \in \mathcal{M}\}$, i.e., $\mathrm{geomed}\left(\{\boldsymbol{z}_m^t | m \in \mathcal{M}\}\right)$. The $\mathrm{geomed}\left(\{\boldsymbol{z}_m^t | m \in \mathcal{M}\}\right)$ in this paper is defined as

$$\mathrm{geomed}\left(\{\boldsymbol{z}_m^t | m \in \mathcal{M}\}\right) \triangleq \boldsymbol{z}^{t,*} = \arg\min_y \sum_{i=1}^{M} \frac{S_i}{\sum_{j=1}^{M} S_j} \left\| y - \boldsymbol{z}_i^t \right\|. \tag{15}$$

The problem in (15) is actually a convex optimization problem and an ϵ-optimal solution is obtainable by resorting to Weiszfeld algorithm [32]. Hence the $\boldsymbol{z}^t$ fulfills the following condition

$$\sum_{i=1}^{M} \frac{S_i}{\sum_{j=1}^{M} S_j} \left\| \boldsymbol{z}^t - \boldsymbol{z}_i^t \right\| \leqslant \sum_{i=1}^{M} \frac{S_i}{\sum_{j=1}^{M} S_j} \left\| \boldsymbol{z}^{t,*} - \boldsymbol{z}_i^t \right\| + \epsilon, \tag{16}$$

and is also denoted as $\mathrm{geomed}\left(\{\boldsymbol{z}_m^t | m \in \mathcal{M}\}, \epsilon\right)$ in the sequel. Then the central server broadcasts the global model parameter $\boldsymbol{w}_{t+1}$ to all the users for preparing the calculation in $t+1$ round of iteration.

Based on the above description, the RAGA is conceptually illustrated in Fig 1 and summarized in Algorithm 1.

IV. Analysis of Convergence and Robustness

In this section, we provide theoretical analysis of the robustness and convergence of the RAGA under Byzantine attacks under two cases of loss function's convexity.

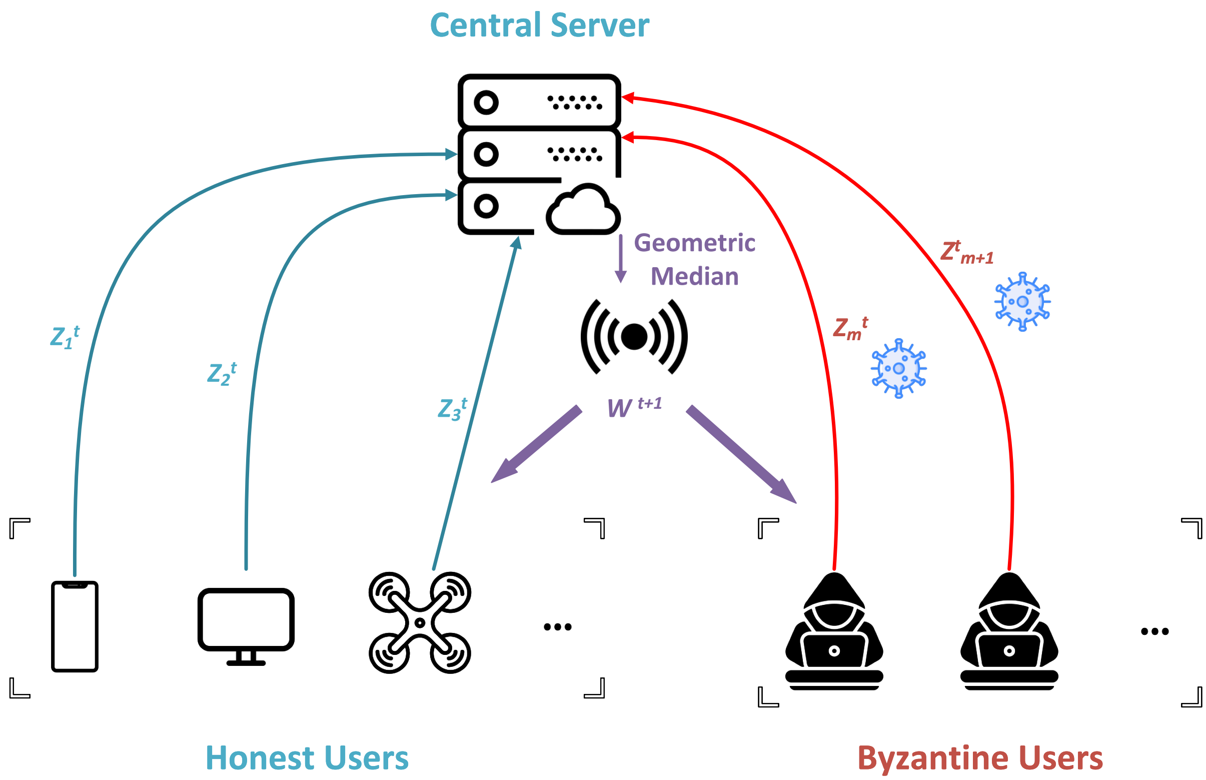

Fig. 1: Illustration of Byzantine-resilient RAGA FL system.

A. *Case I: With Lipschitz Continuity Only for Loss Function*

We first inspect the case with Assumptions 1, 3, 4 and 5 imposed, which allows the loss function to be non-convex. By defining

$$\alpha_m = \frac{S_m}{\sum_{j=1}^{M} S_j}, C_\alpha = \sum_{m \in \mathcal{M} \setminus \mathcal{B}} \alpha_m, \tag{17}$$

and

$$p^t = \begin{cases} 0, & \eta^t \leqslant \dfrac{1}{L} \\ 1, & \eta^t > \dfrac{1}{L} \end{cases} \tag{18}$$

the following results can be expected.

Algorithm 1: The procedure of RAGA.

The central server initializes the model parameter $\boldsymbol{w}^1$ and sends it to all the users.
for $t = 1, 2, \ldots, T$ **do**
  for $m \in \mathcal{M} \setminus \mathcal{B}$ **do**
    Set $\boldsymbol{w}_m^{t,0} = \boldsymbol{w}^t$.
    for $k = 1, 2, \ldots, K^t$ **do**
      Randomly select local data subset $\xi_m^{t,k}$ from $\mathcal{S}_m$ and set
      $\boldsymbol{w}_m^{t,k} = \boldsymbol{w}_m^{t,k-1} - \eta_m^{t,k} \nabla F_m(\boldsymbol{w}_m^{t,k-1}; \xi_m^{t,k})$.
    end
    Set $\boldsymbol{z}_m^t = \frac{1}{K^t} \sum_{k=1}^{K^t} \nabla F_m(\boldsymbol{w}_m^{t,k-1}; \xi_m^{t,k})$.
  end
  for $m \in \mathcal{B}$ **do**
    Set $z_m^t = \star$.
  end
  All the users send their own $\boldsymbol{z}_m^t$ to the central server. The central server obtains the geometric median $\boldsymbol{z}^t = \mathrm{geomed}\left(\{\boldsymbol{z}_m^t | m \in \mathcal{M}\}, \epsilon\right)$, updates $\boldsymbol{w}^{t+1}$ by
  $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta^t \cdot \boldsymbol{z}^t$, and then broadcasts $\boldsymbol{w}^{t+1}$ to all the users.
end

Theorem 1: With $C_\alpha > 0.5$, there is

$$\sum_{t=1}^{T} \frac{\eta^t}{\sum_{t'=1}^{T} \eta^{t'}} \mathbb{E}\left\{ \left\| \nabla F(\boldsymbol{w}^t) \right\|^2 \right\} \leqslant \frac{2\mathbb{E}\left\{ F(\boldsymbol{w}^1) - F(\boldsymbol{w}^T) \right\}}{\sum_{t'=1}^{T} \eta^{t'}} + \frac{1}{\sum_{t'=1}^{T} \eta^{t'}} \sum_{t=1}^{T} \eta^t \Delta^t$$
$$+ \sum_{t=1}^{T} \frac{2(\eta^t + p^t(\eta^t)^2 L - p^t \eta^t)\epsilon^2}{(2C_\alpha - 1)^2 \sum_{t'=1}^{T} \eta^{t'}} + \sum_{t=1}^{T} \frac{8(C_\alpha)^2(\eta^t \sigma^2 + (p^t(\eta^t)^2 L - p^t \eta^t)(G^2 + \sigma^2))}{(2C_\alpha - 1)^2 \sum_{t'=1}^{T} \eta^{t'}} \tag{19}$$

where

$$\Delta^t = \frac{8C_\alpha}{(2C_\alpha - 1)^2} \sum_{m \in \mathcal{M} \setminus \mathcal{B}} \left(\frac{2\alpha_m L^2 (G^2 + \sigma^2)}{K^t} \sum_{k=2}^{K^t} (k-1) \cdot \sum_{i=1}^{k-1} (\eta_m^{t,i})^2 + \frac{2\alpha_m}{K^t} \sum_{k=1}^{K^t} \theta_m^{t,k-1} \right), \tag{20}$$

and

$$\theta_m^{t,k} \triangleq \left\| \nabla F_m(\boldsymbol{w}_m^{t,k}) - \nabla F(\boldsymbol{w}_m^{t,k}) \right\|^2, \tag{21}$$

which actually represents the data heterogeneity of kth update in t iteration for user m.

Proof: Please refer to Appendix A. ■

Remark 1:

- In general case, when $\lim\limits_{T\to\infty}\sum\limits_{t=1}^{T}\eta^t=\infty$, $\lim\limits_{T\to\infty}\sum\limits_{t=1}^{T}\eta^t\sum\limits_{m\in\mathcal{M}\backslash\mathcal{B}}\sum\limits_{i=1}^{K^t}(\eta_m^{t,i})^2<\infty$, and there are only finitely many t such that $p^t=1$, together with the prerequisite condition of Theorem 1, i.e., $C_\alpha>0.5$, it can be derived that the term $\sum\limits_{t=1}^{T}\frac{\eta^t}{\sum_{t'=1}^{T}\eta^t}\mathbb{E}\left\{\left\|\nabla F(\boldsymbol{w}^t)\right\|^2\right\}$ will converge to $\mathcal{O}\left(\max\{\theta_m^{t,k}\}+\epsilon^2+\sigma^2\right)$.
- To be concrete, we set $\eta_m^{t,k}=\eta^t\leqslant\frac{1}{L}$, $K^t=K$, and $\theta_m^{t,k}\leqslant\theta$. Then inequality (19) dwells into

$$\sum_{t=1}^{T}\frac{\eta^t}{\sum_{t'=1}^{T}\eta^t}\mathbb{E}\left\{\left\|\nabla F(\boldsymbol{w}^t)\right\|^2\right\}\leqslant\frac{8(C_\alpha)^2L^2(G^2+\sigma^2)(K-1)(2K-1)\sum_{t=1}^{T}(\eta^t)^3}{3(2C_\alpha-1)^2\sum_{t=1}^{T}\eta^t}$$
$$+\frac{2\mathbb{E}\left\{F(\boldsymbol{w}^1)-F(\boldsymbol{w}^T)\right\}}{\sum_{t=1}^{T}\eta^t}+\frac{16(C_\alpha)^2\theta+2\epsilon^2+8(C_\alpha)^2\sigma^2}{(2C_\alpha-1)^2} \tag{22}$$

  which will converge to

$$\frac{16(C_\alpha)^2}{(2C_\alpha-1)^2}\theta+\frac{2}{(2C_\alpha-1)^2}\epsilon^2+\frac{8(C_\alpha)^2}{(2C_\alpha-1)^2}\sigma^2 \tag{23}$$

  as long as $\lim\limits_{T\to\infty}\sum\limits_{t=1}^{T}(\eta^t)^3<\infty$.
- To further concrete convergence rate, we set $\eta^t=\frac{1}{L}\frac{1}{t^{1/3+\delta}}$, where δ is an arbitrary value lying between $(0,2/3)$, then the right-hand side of (19) is upper bounded by

$$\mathcal{O}\left(\frac{8(C_\alpha)^2(G^2+\sigma^2)(K-1)(2K-1)(1+3\delta-3\delta T^{-3\delta})}{3(2C_\alpha-1)^2T^{2/3-\delta}}\right)$$
$$+\mathcal{O}\left(\frac{2L\mathbb{E}\left\{F(\boldsymbol{w}^1)-F(\boldsymbol{w}^T)\right\}}{T^{2/3-\delta}}\right)+\frac{16(C_\alpha)^2\theta+2\epsilon^2+8(C_\alpha)^2\sigma^2}{(2C_\alpha-1)^2} \tag{24}$$

  which still converges to the result in (23) at rate $\mathcal{O}\left(\frac{1}{T^{2/3-\delta}}\right)$.
- From previous inference from Theorem 1, it can be found that the minimum of $\mathbb{E}\left\{\left\|\nabla F(\boldsymbol{w}^t)\right\|^2\right\}$ over $t\in[1,T]$, which is upper bounded by the left-hand side of (19), and thus the expression in (23). The non-zero gap shown in (23) will vanish as θ, ϵ and σ, which represent data heterogeneity, error tolerance to numerically work out the geometric median through Weiszfeld algorithm, and the variance of stochastic local gradient respectively, go to zero. In other words, once the $\mathcal{S}_m$ among $m\in\mathcal{M}$ are IID, $\mathcal{S}_m$ for every $m\in\mathcal{M}$ are

fully utilized to calculate local gradient, and the geometric median is calculated perfectly, there will be one $\nabla F(\boldsymbol{w}^t)$ for $t \in [1, T]$ being zero, which implies the reaching of stationary point for Problem 1.

- Last but not least, the above convergence results are based on the assumption that $C_\alpha > 0.5$, i.e., the fraction of the datasets from Byzantine attackers is less than half, which shows a strong robustness to Byzantine attacks.

B. Case II: With Lipschitz Continuity and Strong Convexity for Loss Function

For the case with Assumptions 1, 2, 3, 4 and 5 imposed, which requires the loss function to be not only Lipschitz but also strongly-convex, the following results can be anticipated.

Theorem 2: With $0 < \lambda^t < 1$ and $C_\alpha > 0.5$, the optimality gap $\mathbb{E}\left\{F(\boldsymbol{w}^T) - F(\boldsymbol{w}^*)\right\}$ is upper bounded in (25) as follows,

$$\begin{aligned} &\mathbb{E}\left\{F(\boldsymbol{w}^T) - F(\boldsymbol{w}^*)\right\} \\ &\leqslant \frac{L}{2}\mathbb{E}\left\{\left\|\boldsymbol{w}^1 - \boldsymbol{w}^*\right\|^2\right\}\prod_{t=1}^{T-1}\gamma^t + \frac{L}{2}\sum_{t=1}^{T-1}\frac{(\eta^t)^2}{\lambda^t}\left(\Delta^t + \frac{8\sigma^2(C_\alpha)^2}{(2C_\alpha-1)^2} + \frac{2\epsilon^2}{\left(2C_\alpha-1\right)^2}\right)\cdot\prod_{i=t}^{T-1}(\gamma^{i+1})^{q^{i+1}} \end{aligned} \tag{25}$$

where

$$q^t = \begin{cases} 0, & t = T \\ 1, & t < T \end{cases}$$

and

$$\gamma^t = \frac{1 - 2\eta^t\mu + L^2(\eta^t)^2}{1 - \lambda^t}.$$

Proof: Please refer to Appendix B. ∎

Remark 2:

- For general case, as long as $\eta_m^{t,k} \propto T^{-\beta}, \beta > 0$, $\lambda^t = \rho^t\eta^t$ with $\rho^t < 2\mu$ and $\eta^t < \dfrac{2\mu - \rho^t}{L^2}$, $0 < \gamma^t < 1$, and the preconditions listed in Theorem 2 hold, the convergence of our proposed RAGA can be ensured and the optimality gap is at the scale $\mathcal{O}\left(\max\{\theta_m^{t,k}\} + \epsilon^2 + \sigma^2\right)$.

- To concrete the optimality gap, we suppose $\eta_m^{t,k} = \frac{1}{T}\eta$, $\eta^t = \eta$, $\gamma^t = \gamma$, $K^t = K$, $\lambda^t = \mu\eta^t = \mu\eta$, $\eta^t < \frac{\mu}{L^2}$, and $\theta_m^{t,k} \leqslant \theta$, then there is

$$\begin{aligned}&\mathbb{E}\left\{F(\boldsymbol{w}^T) - F(\boldsymbol{w}^*)\right\} \leqslant \frac{L(\gamma)^{T-1}}{2}\mathbb{E}\left\{\left\|\boldsymbol{w}^1 - \boldsymbol{w}^*\right\|^2\right\}\\ &+ \frac{4L^3(K-1)(2K-1)(C_\alpha)^2(G^2+\sigma^2)}{3\mu(2C_\alpha - 1)^2T^2}\frac{1-(\gamma)^T}{1-\gamma}(\eta)^3 + \frac{8L(C_\alpha)^2\theta + \epsilon^2 L + 4L(C_\alpha)^2\sigma^2}{\mu(2C_\alpha-1)^2}\frac{1-(\gamma)^T}{1-\gamma}\eta.\end{aligned} \tag{26}$$

When $T \to \infty$, $(\gamma)^{T-1} \to 0$ and $\frac{1-(\gamma)^T}{1-\gamma} \to \frac{1}{1-\gamma} = \frac{1-\mu\eta}{\mu\eta - L^2(\eta)^2} < \infty$. Therefore, $\frac{1-(\gamma)^T}{T^2(1-\gamma)} \to \frac{1-\mu\eta}{T^2(\mu\eta - L^2(\eta)^2)} \to 0$ and $\frac{\eta(1-(\gamma)^T)}{1-\gamma} \to \frac{1-\mu\eta}{\mu - L^2\eta}$. In this case, $\mathbb{E}\left\{F(\boldsymbol{w}^T) - F(\boldsymbol{w}^*)\right\}$ will converge to

$$\frac{1-\mu\eta}{\mu - L^2\eta}\left(\frac{8L(C_\alpha)^2}{\mu(2C_\alpha-1)^2}\theta + \frac{L}{\mu(2C_\alpha-1)^2}\epsilon^2 + \frac{4L(C_\alpha)^2}{\mu(2C_\alpha-1)^2}\sigma^2\right) \tag{27}$$

at a rate $\mathcal{O}\left((\gamma)^{T-1} + \frac{1}{T^2}\right)$.

- To further concrete convergence rate, we alternatively set $\eta_m^{t,k} = (TK)^{-\beta}\eta$ with $\beta > 0$, then the optimality gap of (25) can be bounded by

$$\begin{aligned}&\left(\frac{4L^3(K-1)(2K-1)(C_\alpha)^2(G^2+\sigma^2)}{3\mu(2C_\alpha-1)^2(TK)^{2\beta}}\frac{1-(\gamma)^T}{1-\gamma}(\eta)^3\right)\\ &+\left(\frac{L(\gamma)^{T-1}}{2}\mathbb{E}\left\{\left\|\boldsymbol{w}^1 - \boldsymbol{w}^*\right\|^2\right\}\right) + \frac{8L(C_\alpha)^2\theta + \epsilon^2 L + 4L(C_\alpha)^2\sigma^2}{\mu(2C_\alpha-1)^2}\frac{1-(\gamma)^T}{1-\gamma}\eta\end{aligned} \tag{28}$$

which also convergences to (27) but at a rate $\mathcal{O}\left((\gamma)^{T-1} + (TK)^{-2\beta}\right)$. This rate can be adjusted by β and would be linear when $\beta \propto T$.

- Compared with the discussion in Remark 1 that works for non-convex loss function, we can obtain zero optimality gap as θ, ϵ, and σ go to zero for strongly-convex loss function in this case, which implies the achievement of global optimal solution rather than stationary point as in Remark 1. Moreover, linear convergence rate is achievable in this case and is faster than the rate $\mathcal{O}\left(\frac{1}{T^{2/3-\delta}}\right)$ as shown in Remark 1.
- It is also worthy to note that the above results still hold on the condition that $C_\alpha > 0.5$, which promises a strong robustness to Byzantine attacks.

V. Experiments

In this section, we conduct experiments to inspect the convergence performance of our proposed RAGA under various system setups together with baseline methods, so as to verify the advantage of RAGA over peers in being robust to Byzantine attacks in front of heterogeneous dataset.

TABLE II: Test accuracy of RAGA, FedAvg, Byrd-SAGA, and RFA with five intensity levels of Byzantine attack on LeNet and MLP learning models. The iteration number T of CIFAR10 and MNIST are 10000 and 1000 respectively.

Dataset	**CIFAR10**										**MNIST**									
Model	**LeNet**					**MLP**					**LeNet**					**MLP**				
Level of Byzantine Attacks $(1-C_\alpha)$	0	0.1	0.2	0.3	0.4	0	0.1	0.2	0.3	0.4	0	0.1	0.2	0.3	0.4	0	0.1	0.2	0.3	0.4
RAGA (%)	79.25	**79.09**	**78.18**	**77.13**	**77.14**	58.46	**59.04**	57.74	**58.17**	**56.59**	96.53	**96.57**	**96.50**	**96.39**	**96.27**	92.90	**92.91**	**92.77**	**92.84**	**92.69**
FedAvg (%)	**80.02**	10.00	10.00	10.00	10.00	**59.76**	10.00	10.00	10.00	10.00	**96.92**	11.35	11.35	11.35	10.32	**93.07**	11.35	10.28	10.28	9.58
Byrd-SAGA (%)	72.13	72.35	70.46	72.33	71.98	56.10	55.94	55.57	55.85	55.50	93.01	93.06	92.91	92.57	92.90	90.99	90.91	90.93	90.69	90.78
RFA (%)	78.07	77.42	77.14	76.60	75.52	58.86	58.18	**57.79**	57.59	56.44	96.01	95.96	95.91	95.87	95.86	92.49	92.48	92.52	92.35	92.34

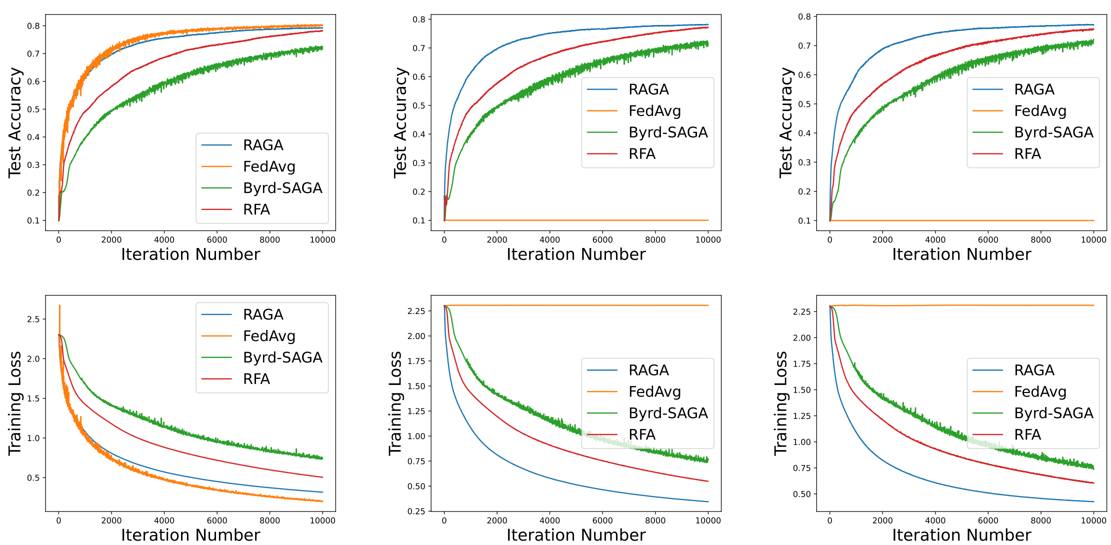

Fig. 2: Text accuracy and training loss of LeNet with CIFAR10 dataset for RAGA, FedAvg, Byrd-SAGA, and RFA. The intensity levels of Byzantine attack $(1-C_\alpha)$ are 0, 0.2, 0.4 from left to right.

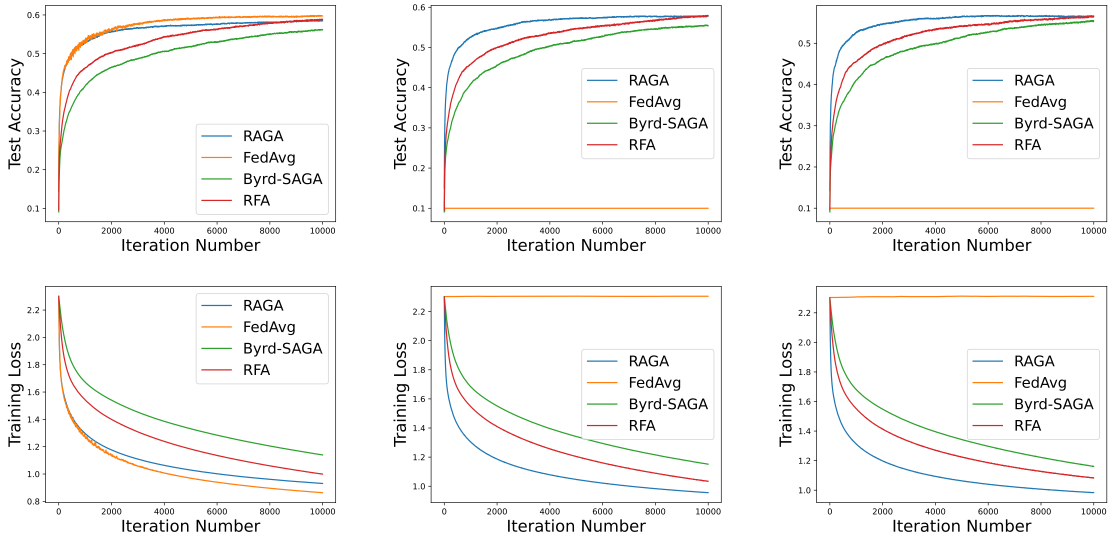

Fig. 3: Text accuracy and training loss of MLP with CIFAR10 dataset for RAGA, FedAvg, Byrd-SAGA, and RFA. The intensity levels of Byzantine attack $(1 - C_\alpha)$ are 0, 0.2, 0.4 from left to right.

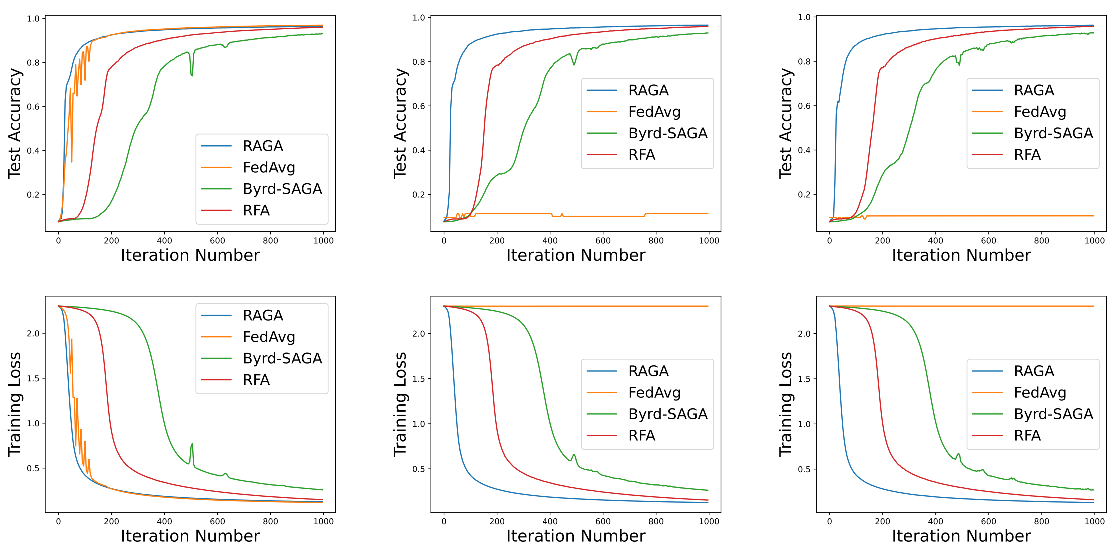

Fig. 4: Text accuracy and training loss of LeNet with MNIST dataset for RAGA, FedAvg, Byrd-SAGA, and RFA. The intensity levels of Byzantine attack $(1 - C_\alpha)$ are 0, 0.2, 0.4 from left to right.

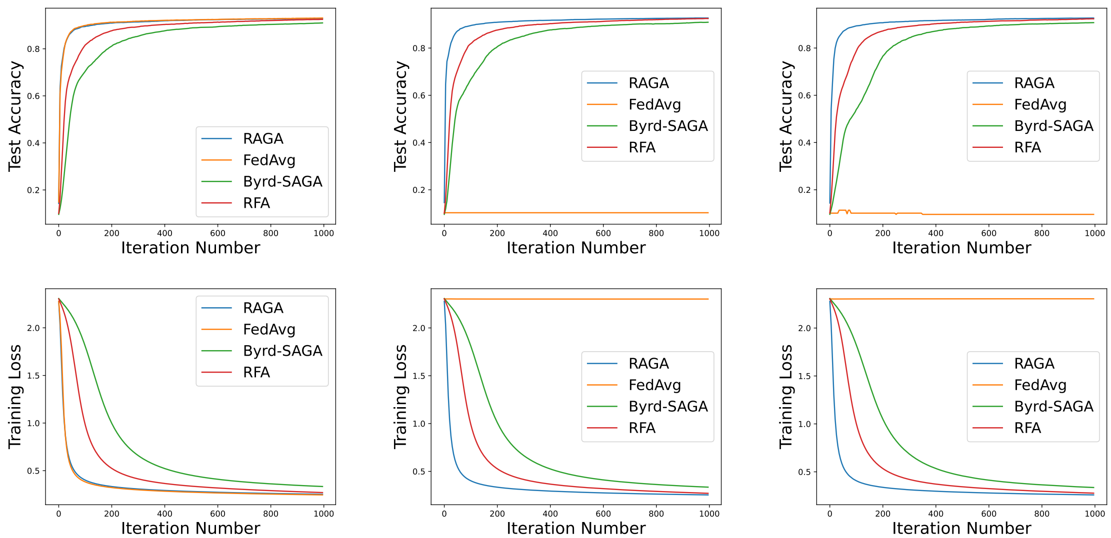

Fig. 5: Text accuracy and training loss of MLP with MNIST dataset for RAGA, FedAvg, Byrd-SAGA, and RFA. The intensity levels of Byzantine attack $(1 - C_\alpha)$ are 0, 0.2, 0.4 from left to right.

A. *Setup*

To carry out experiments, we set up a machine learning environment in PyTorch 2.1.0 on Ubuntu 20.04, powered by two 3090 GPUs.

1) Datasets and Models: We conduct experiments by leveraging two datasets: MNIST and CIFAR10. These two datasets are widely used for learning tasks and are introduced as follows:

- **MNIST:** The MNIST dataset includes a training set and a test set. The training set contains 60000 samples and the test set contains 10000 samples, each sample of which is a 28 $\times$ 28 pixel grayscale image.
- **CIFAR10:** The CIFAR10 dataset is composed of a training set and a test set. The training set contains 50000 samples and the test set contains 10000 samples, where each sample is a 32 $\times$ 32 color image.

For these experiments, we split the above two training sets into M non-IID training sets, which is realized by letting the label of data samples to conform to Dirichlet distribution. The extent of non-IID can be adjusted by tuning the concentration parameter of Dirichlet distribution.

For training problems with non-convex and strongly-convex loss functions, we adopt LeNet and Multilayer Perceptron (MLP) model, respectively. The introduction of these two models are as follows:

- **LeNet:** The LeNet model is one of the earliest published convolutional neural networks. For the experiments, we are going to train a LeNet model with two convolutional layers, two pooling layers, and one fully connected layer like in [33]. Cross-entropy function is taken as the training loss.
- **MLP:** The MLP model is a machine learning model based on Feedforward Neural Network that can achieve high-level feature extraction and classification. To promise strong convexity for the loss function, we configure the MLP model to be with only fully connected layers like in [34] and take the cross-entropy function as the training loss.

2) Hyperparameters: We assume $M = 50$ and the batchsize of all the experiments is 512. The error tolerance ϵ for numerically working out geometric median is $\epsilon = 1 \times 10^{-5}$. The concentration parameter of Dirichlet distribution is 0.6. In default, the iteration number $T = 1000$ for the MNIST dataset, and $T = 10000$ for the CIFAR10 dataset. The round number of local updates K^t is 3 for RFA and RAGA. The Byzantine attacks are selected as the Gaussian attack, which obeys standard normal distribution. $1 - C_\alpha$ are used to indicate the intensity level of Byzantine attacks. In terms of the learning rate, we adopt default setup as shown in corresponding papers for baseline methods. For RAGA, the learning rate $\eta_m^{t,k} = \eta^t = T/(100t + 10T)$.

3) Baselines: The convergence performance of four algorithms: RAGA, FedAvg [17], Byrd-SAGA [26], and RFA [25] are compared. For the latter three algorithms, FedAvg is renowned in traditional FL but does not concern Byzantine attack and can be taken as a performance metric when there is no Byzantine attack, Byrd-SAGA only executes one local updating in each round of iteration, and RFA selects the trimmed mean of the model parameters over multiple local updating as the uploaded vector. It is worthy to note that there is another Byzantine-resilient FL algorithm in related works: BROADCAST [27]. We do not show the performance of it because BROADCAST is very similar with Byrd-SAGA and only differs from Byrd-SAGA by uploading vector from each user to the central server in a compressive, rather than lossless way like in Byrd-SAGA.

4) Performance Metrics: To assess the convergence performance, we measure the test accuracy, the convergence speed and the training loss. An algorithm would be better if its test accuracy is higher, convergence speed is faster, or training loss is lower.

B. Convergence Performance

We start from Table II, which shows the test accuracy of these four algorithms under different intensity levels of Byzantine attack for various datasets and learning models. From Table II, we see that RAGA can mostly achieve the highest test accuracy under Byzantine attacks. Even there is no Byzantine attack, in which case FedAvg performs best, RAGA ranks the second in most cases and is very close to FedAvg. Additionally, as the intensity level of Byzantine attack grows from 0 to 0.4, in which case nearly half of training data of the whole system is corrupted by Byzantine users, we see RAGA degrades little, which verifies the robustness of RAGA to Byzantine attacks.

Fig. 2 and Fig. 3 illustrate the convergence performance of the four comparing algorithms versus iteration number, by means of test accuracy and training loss, in the training task of LeNet and MLP respectively, for CIFAR10 dataset. From Fig. 2 and Fig. 3, we can observe that both RAGA and FedAvg act similarly and perform better than other algorithms when there is no Byzantine attack. With the existence of Byzantine attack, FedAvg performs badly as it does not take Byzantine attack into account and RAGA becomes the one that outperforms all the other algorithms, both in terms of test accuracy and training loss.

Fig. 4 and Fig. 5 repeat the experiment as shown in Fig. 2 and Fig. 3 but for a different dataset, i.e., MNIST dataset. From Fig. 4 and Fig. 5, when there is no Byzantine attack, we can see that both RAGA and FedAvg have the best convergence performance but RAGA is more stable than FedAvg for the training task of LeNet model. As Byzantine attack comes up, FedAvg is very sensitive to the Byzantine attack and degrades a lot. In contrast, RAGA is always the best one compared with all the other algorithms, no matter how strong the Byzantine attack is.

In summary, through the above numerical experiments, it can be seen that the RAGA not only defeats its opponents at different levels of Byzantine attack, but also acts well and has a comparable performance with FedAvg that performs best when there is no Byzantine attack. It is also worthy to mention that RAGA achieves a nearly unchanged convergence performance no matter how strong the Byzantine attack is. Last but not least, all the above results are obtained based on non-IID datasets.

VI. Conclusion

In this paper, we propose a novel FL aggregation strategy, RAGA, so as to be robust to Byzantine attacks and data heterogeneity. In our proposed RAGA, geometric median is leveraged for aggregation and the round number for local updating can be selected freely. Rigorous proofs disclose that our proposed RAGA can achieve convergence at rate $\mathcal{O}(1/T^{2/3-\delta})$ with $\delta \in (0, 2/3)$ for non-convex loss function and at linear rate for strongly-convex loss function, so long as the fraction of dataset from malicious users is less than half. Experimental results based on heterogeneous datasets show that RAGA performs steadily and keeps on an advantage on convergence performance over baseline methods under various intensity of Byzantine attacks.

APPENDIX A
PROOF OF THEOREM 1

The proof of Theorem 1 relies on the holding of Lemma 1 and Lemma 2, which is given as follows:

Lemma 1: With Assumptions 1, 3, 4, 5, and $C_\alpha > 0.5$, the term $\mathbb{E}\left\{\left\|\boldsymbol{z}^t - \nabla F(\boldsymbol{w}^t)\right\|^2\right\}$ can be upper bounded as

$$\mathbb{E}\left\{\left\|\boldsymbol{z}^t - \nabla F(\boldsymbol{w}^t)\right\|^2\right\} \leqslant \Delta^t + \frac{8\sigma^2(C_\alpha)^2}{(2C_\alpha - 1)^2} + \frac{2\epsilon^2}{(2C_\alpha - 1)^2} \tag{29}$$

Proof: Please refer to Appendix C. ■

Lemma 2: With Assumptions 3, 4, 5, and $C_\alpha > 0.5$, the term $\mathbb{E}\left\{\left\|\boldsymbol{z}^t\right\|^2\right\}$ can be bounded as

$$0 \leqslant \mathbb{E}\left\{\left\|\boldsymbol{z}^t\right\|^2\right\} \leqslant \frac{8(C_\alpha)^2(G^2+\sigma^2)}{(2C_\alpha - 1)^2} + \frac{2\epsilon^2}{(2C_\alpha - 1)^2} \tag{30}$$

Proof: Please refer to Appendix D. ■

Under Assumption 1, Lemmas 1 and 2, and recalling the definition of p^t as given in (18), we have

$$\begin{aligned}
&\mathbb{E}\left\{F(\boldsymbol{w}^{t+1}) - F(\boldsymbol{w}^t)\right\} \\
&\leqslant \mathbb{E}\left\{\left\langle \nabla F(\boldsymbol{w}^t), \boldsymbol{w}^{t+1} - \boldsymbol{w}^t \right\rangle\right\} + \frac{L}{2}\mathbb{E}\left\{\left\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\right\|^2\right\} && \text{(31a)} \\
&= -\eta^t \mathbb{E}\left\{\left\langle \nabla F(\boldsymbol{w}^t), \boldsymbol{z}^t \right\rangle\right\} + \frac{(\eta^t)^2 L}{2}\mathbb{E}\left\{\left\|\boldsymbol{z}^t\right\|^2\right\} && \text{(31b)} \\
&= \frac{\eta^t}{2}\mathbb{E}\left\{\left\|\boldsymbol{z}^t - \nabla F(\boldsymbol{w}^t)\right\|^2\right\} - \frac{\eta^t}{2}\mathbb{E}\left\{\left\|\nabla F(\boldsymbol{w}^t)\right\|^2\right\} + \frac{(\eta^t)^2 L - \eta^t}{2}\mathbb{E}\left\{\left\|\boldsymbol{z}^t\right\|^2\right\} && \text{(31c)} \\
&= \frac{\eta^t}{2}\mathbb{E}\left\{\left\|\boldsymbol{z}^t - \nabla F(\boldsymbol{w}^t)\right\|^2\right\} - \frac{\eta^t}{2}\mathbb{E}\left\{\left\|\nabla F(\boldsymbol{w}^t)\right\|^2\right\} + \frac{(\eta^t)^2 L - \eta^t}{2}p^t\mathbb{E}\left\{\left\|\boldsymbol{z}^t\right\|^2\right\} && \text{(31d)} \\
&\leqslant -\frac{\eta^t}{2}\mathbb{E}\left\{\left\|\nabla F(\boldsymbol{w}^t)\right\|^2\right\} + \frac{\eta^t}{2}\Delta^t + \frac{(\eta^t + p^t(\eta^t)^2 L - p^t\eta^t)\epsilon^2}{(2C_\alpha - 1)^2} + \frac{4(C_\alpha)^2(\eta^t\sigma^2 + (p^t(\eta^t)^2 L - p^t\eta^t)(G^2+\sigma^2))}{(2C_\alpha - 1)^2} && \text{(31e)}
\end{aligned}$$

Summarizing the inequality in (31) for $t = 1, 2, ..., T$ and dividing the summarized inequality with $\sum_{t=1}^{T} \eta^t / 2$ for both sides, we obtain

$$
\begin{aligned}
&\sum_{t=1}^{T} \frac{\eta^t}{\sum_{t'=1}^{T} \eta^t} \mathbb{E}\left\{ \left\| \nabla F(\boldsymbol{w}^t) \right\|^2 \right\} \\
&\leqslant \frac{2\mathbb{E}\left\{ F(\boldsymbol{w}^1) - F(\boldsymbol{w}^T) \right\}}{\sum_{t'=1}^{T} \eta^t} + \frac{1}{\sum_{t=1}^{T} \eta^t} \sum_{t=1}^{T} \eta^t \Delta^t \\
&\quad + \sum_{t=1}^{T} \frac{2(\eta^t + p^t(\eta^t)^2 L - p^t \eta^t)\epsilon^2}{(2C_\alpha - 1)^2 \sum_{t'=1}^{T} \eta^t} + \sum_{t=1}^{T} \frac{8(C_\alpha)^2(\eta^t \sigma^2 + (p^t(\eta^t)^2 L - p^t \eta^t)(G^2 + \sigma^2))}{(2C_\alpha - 1)^2 \sum_{t'=1}^{T} \eta^t}
\end{aligned} \tag{32}
$$

This completes the proof of Theorem 1.

Appendix B
Proof of Theorem 2

With Assumptions 1 and the fact that $\nabla F(\boldsymbol{w}^*) = \boldsymbol{0}$, there is

$$
\begin{aligned}
&\mathbb{E}\left\{ F(\boldsymbol{w}^{t+1}) - F(\boldsymbol{w}^*) \right\} \\
&\leqslant \mathbb{E}\left\{ \left\langle \nabla F(\boldsymbol{w}^*), \boldsymbol{w}^{t+1} - \boldsymbol{w}^* \right\rangle \right\} + \frac{L}{2} \mathbb{E}\left\{ \left\| \boldsymbol{w}^{t+1} - \boldsymbol{w}^* \right\|^2 \right\} && \text{(33a)} \\
&= \frac{L}{2} \mathbb{E}\left\{ \left\| \boldsymbol{w}^{t+1} - \boldsymbol{w}^* \right\|^2 \right\} && \text{(33b)}
\end{aligned}
$$

For the term $\mathbb{E}\left\{ \left\| \boldsymbol{w}^{t+1} - \boldsymbol{w}^* \right\|^2 \right\}$, with $0 < \lambda^t < 1$, it leads to

$$
\begin{aligned}
&\mathbb{E}\left\{ \left\| \boldsymbol{w}^{t+1} - \boldsymbol{w}^* \right\|^2 \right\} \\
&= \mathbb{E}\left\{ \left\| \boldsymbol{w}^{t+1} - \boldsymbol{w}^t + \eta^t \nabla F(\boldsymbol{w}^t) + \boldsymbol{w}^t - \eta^t \nabla F(\boldsymbol{w}^t) - \boldsymbol{w}^* \right\|^2 \right\} && \text{(34a)} \\
&\leqslant \frac{1}{\lambda^t} \mathbb{E}\left\{ \left\| \boldsymbol{w}^{t+1} - \boldsymbol{w}^t + \eta^t \nabla F(\boldsymbol{w}^t) \right\|^2 \right\} + \frac{1}{1 - \lambda^t} \mathbb{E}\left\{ \left\| \boldsymbol{w}^t - \eta^t \nabla F(\boldsymbol{w}^t) - \boldsymbol{w}^* \right\|^2 \right\} && \text{(34b)} \\
&= \frac{(\eta^t)^2}{\lambda^t} \mathbb{E}\left\{ \left\| \boldsymbol{z}^t - \nabla F(\boldsymbol{w}^t) \right\|^2 \right\} + \frac{1}{1 - \lambda^t} \mathbb{E}\left\{ \left\| \boldsymbol{w}^t - \boldsymbol{w}^* \right\|^2 \right\} - \frac{2\eta^t}{1 - \lambda^t} \left\langle \boldsymbol{w}^t - \boldsymbol{w}^*, \nabla F(\boldsymbol{w}^t) - \nabla F(\boldsymbol{w}^*) \right\rangle \\
&\quad + \frac{(\eta^t)^2}{1 - \lambda^t} \mathbb{E}\left\{ \left\| \nabla F(\boldsymbol{w}^t) - \nabla F(\boldsymbol{w}^*) \right\|^2 \right\} && \text{(34c)} \\
&\leqslant \frac{(\eta^t)^2}{\lambda^t} \mathbb{E}\left\{ \left\| \boldsymbol{z}^t - \nabla F(\boldsymbol{w}^t) \right\|^2 \right\} + \frac{1 - 2\mu\eta^t + L^2(\eta^t)^2}{1 - \lambda^t} \mathbb{E}\left\{ \left\| \boldsymbol{w}^t - \boldsymbol{w}^* \right\|^2 \right\} && \text{(34d)}
\end{aligned}
$$

where

- the inequality (34b) comes from Cauchy-Schwarz inequality $\left(\frac{a^2}{c} + \frac{b^2}{1 - c} \right)(c + 1 - c) \geqslant (a + b)^2, 0 < c < 1$;

- the equality (34c) is established because of $\nabla F(\boldsymbol{w}^*) = \mathbf{0}$;
- the inequality (34d) is bounded by Assumptions 1 and 2.

Applying the inequality in (34) for $t = 1, 2, ..., T$, and recalling that $\gamma^t = \dfrac{1 - 2\mu\eta^t + L^2(\eta^t)^2}{1 - \lambda^t}$ and the definition of q^t, we then have

$$\mathbb{E}\left\{\left\|\boldsymbol{w}^T - \boldsymbol{w}^*\right\|^2\right\} \leqslant \prod_{t=1}^{T-1} \gamma^t \mathbb{E}\left\{\left\|\boldsymbol{w}^1 - \boldsymbol{w}^*\right\|^2\right\} + \sum_{t=1}^{T-1} \frac{(\eta^t)^2}{\lambda^t} \prod_{i=t}^{T-1} (\gamma^{i+1})^{q^{i+1}} \mathbb{E}\left\{\left\|\boldsymbol{z}^t - \nabla F(\boldsymbol{w}^t)\right\|^2\right\} \tag{35}$$

With the help of Lemma 1, and combining (33) and (35), there is

$$\begin{aligned} &\mathbb{E}\left\{F(\boldsymbol{w}^T) - F(\boldsymbol{w}^*)\right\} \\ &\leqslant \frac{L}{2}\mathbb{E}\left\{\left\|\boldsymbol{w}^1 - \boldsymbol{w}^*\right\|^2\right\} \prod_{t=1}^{T-1} \gamma^t + \frac{L}{2} \sum_{t=1}^{T-1} \frac{(\eta^t)^2}{\lambda^t} \left[\Delta^t + \frac{8\sigma^2 (C_\alpha)^2}{(2C_\alpha - 1)^2} + \frac{2\epsilon^2}{(2C_\alpha - 1)^2}\right] \cdot \prod_{i=t}^{T-1} (\gamma^{i+1})^{q^{i+1}} \end{aligned} \tag{36}$$

This completes the proof of Theorem 2.

Appendix C
Proof of Lemma 1

The proof of Lemma 1 relies on Lemma 3, i.e.,

Lemma 3: Let $\mathcal{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_M\}$ be a set of uploaded vectors by M users. The $\mathcal{Z}$ contains B Byzantine attack vectors, which compose of a subset $\mathcal{Z}'$. When $C_\alpha > 0.5$, there is

$$\begin{aligned} &\mathbb{E}\left\{\|\text{geomed}\left(\{\boldsymbol{z}_m | m \in \mathcal{M}\}, \epsilon\right)\|^2\right\} \\ &\leqslant \frac{8C_\alpha}{(2C_\alpha - 1)^2} \sum_{\boldsymbol{z}_i \in \mathcal{Z} \setminus \mathcal{Z}'} \alpha_i \mathbb{E}\left\|\boldsymbol{z}_i\right\|^2 + \frac{2\epsilon^2}{(2C_\alpha - 1)^2} \end{aligned} \tag{37}$$

Proof: Please refer to Appendix E. ∎

With Lemma 3, there is

$$\begin{aligned} \mathbb{E}\left\{\left\|\boldsymbol{z}^t - \nabla F(\boldsymbol{w}^t)\right\|^2\right\} &\leqslant \frac{2\epsilon^2}{(2C_\alpha - 1)^2} \\ &+ \frac{8C_\alpha}{(2C_\alpha - 1)^2} \sum_{m \in \mathcal{M} \setminus \mathcal{B}} \alpha_m \mathbb{E}\left\|\boldsymbol{z}_m^t - \nabla F(\boldsymbol{w}^t)\right\|^2 \end{aligned} \tag{38}$$

since

$\boldsymbol{z}^t - \nabla F(\boldsymbol{w}^t) = \text{geomed}\left(\{\boldsymbol{z}_m^t - \nabla F(\boldsymbol{w}^t) | m \in \mathcal{M}\}, \epsilon\right)$.

For the term $\mathbb{E}\left\{\left\|\boldsymbol{z}_m^t-\nabla F(\boldsymbol{w}^t)\right\|^2\right\}$, $m \in \mathcal{M}\setminus\mathcal{B}$, we have

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\boldsymbol{z}_m^t-\nabla F(\boldsymbol{w}^t)\right\|^2\right\}\\
&=\mathbb{E}\left\{\left\|\frac{1}{K^t}\sum_{k=1}^{K^t}\nabla F_m(\boldsymbol{w}_m^{t,k-1};\xi_m^{t,k})-\nabla F(\boldsymbol{w}^t)\right\|^2\right\} &\text{(39a)}\\
&=\frac{1}{(K^t)^2}\mathbb{E}\left\{\left\|\sum_{k=1}^{K^t}\left(\nabla F_m(\boldsymbol{w}_m^{t,k-1};\xi_m^{t,k})-\nabla F(\boldsymbol{w}^t)\right)\right\|^2\right\} &\text{(39b)}\\
&\leqslant\frac{1}{K^t}\sum_{k=1}^{K^t}\mathbb{E}\left\{\left\|\nabla F_m(\boldsymbol{w}_m^{t,k-1};\xi_m^{t,k})-\nabla F(\boldsymbol{w}^t)\right\|^2\right\} &\text{(39c)}\\
&=\frac{1}{K^t}\sum_{k=1}^{K^t}\mathbb{E}\Big\{\left\|\nabla F_m(\boldsymbol{w}_m^{t,k-1};\xi_m^{t,k})-\nabla F_m(\boldsymbol{w}_m^{t,k-1})+\nabla F_m(\boldsymbol{w}_m^{t,k-1})-\nabla F(\boldsymbol{w}^t)\right\|^2\Big\} &\text{(39d)}\\
&=\frac{1}{K^t}\sum_{k=1}^{K^t}\mathbb{E}\Big\{\left\|\nabla F_m(\boldsymbol{w}_m^{t,k-1};\xi_m^{t,k})-\nabla F_m(\boldsymbol{w}_m^{t,k-1})\right\|^2+\left\|\nabla F_m(\boldsymbol{w}_m^{t,k-1})-\nabla F(\boldsymbol{w}^t)\right\|^2\\
&\quad+2\left\langle\nabla F_m(\boldsymbol{w}_m^{t,k-1};\xi_m^{t,k})-\nabla F_m(\boldsymbol{w}_m^{t,k-1}),\nabla F_m(\boldsymbol{w}_m^{t,k-1})-\nabla F(\boldsymbol{w}^t)\right\rangle\Big\} &\text{(39e)}
\end{aligned}$$

where the inequality (39c) holds because of the Cauchy-Schwarz inequality $(a_1+a_2+a_3+\cdots+a_n)^2\leqslant n(a_1^2+a_2^2+a_3^2+\cdots+a_n^2)$. With Assumptions 3 and 4, there are $\mathbb{E}\left\{\nabla F_m(\boldsymbol{w}_m^{t,k-1};\xi_m^{t,k})-\nabla F_m(\boldsymbol{w}_m^{t,k-1})\right\}=0$ and $\mathbb{E}\left\{\left\|\nabla F_m(\boldsymbol{w}_m^{t,k-1};\xi_m^{t,k})-\nabla F_m(\boldsymbol{w}_m^{t,k-1})\right\|^2\right\}\leqslant\sigma^2$, then (39) leads to

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\boldsymbol{z}_m^t-\nabla F(\boldsymbol{w}^t)\right\|^2\right\}\\
&\leqslant\frac{1}{K^t}\sum_{k=1}^{K^t}\mathbb{E}\left\{\left\|\nabla F_m(\boldsymbol{w}_m^{t,k-1})-\nabla F(\boldsymbol{w}^t)\right\|^2\right\}+\sigma^2 &\text{(40a)}\\
&=\frac{1}{K^t}\sum_{k=1}^{K^t}\mathbb{E}\Big\{\left\|\nabla F_m(\boldsymbol{w}_m^{t,k-1})-\nabla F(\boldsymbol{w}_m^{t,k-1})+\nabla F(\boldsymbol{w}_m^{t,k-1})-\nabla F(\boldsymbol{w}^t)\right\|^2\Big\}+\sigma^2 &\text{(40b)}\\
&\leqslant\frac{2}{K^t}\sum_{k=1}^{K^t}\mathbb{E}\left\{\left\|\nabla F(\boldsymbol{w}_m^{t,k-1})-\nabla F(\boldsymbol{w}^t)\right\|^2\right\}+\frac{2}{K^t}\sum_{k=1}^{K^t}\mathbb{E}\left\{\left\|\nabla F_m(\boldsymbol{w}_m^{t,k})-\nabla F(\boldsymbol{w}_m^{t,k})\right\|^2\right\}+\sigma^2 &\text{(40c)}
\end{aligned}$$

where the inequality (40c) can be derived from $(a+b)^2\leqslant 2a^2+2b^2$.

To further upper bound the right-hand side of (40), since $\theta_m^{t,k}=\left\|\nabla F_m(\boldsymbol{w}_m^{t,k})-\nabla F(\boldsymbol{w}_m^{t,k})\right\|^2$, we only need to inspect the bound of $\sum_{k=1}^{K^t}\mathbb{E}\left\{\left\|\nabla F(\boldsymbol{w}_m^{t,k-1})-\nabla F(\boldsymbol{w}^t)\right\|^2\right\}$, which is given as follows

$$\sum_{k=1}^{K^t} \mathbb{E}\left\{\left\|\nabla F(\boldsymbol{w}_m^{t,k-1}) - \nabla F(\boldsymbol{w}^t)\right\|^2\right\}$$

$$\leqslant L^2 \sum_{k=1}^{K^t} \mathbb{E}\left\{\left\|\boldsymbol{w}_m^{t,k-1} - \boldsymbol{w}^t\right\|^2\right\} \tag{41a}$$

$$= L^2 \sum_{k=2}^{K^t} \mathbb{E}\left\{\left\|\sum_{i=1}^{k-1} \eta_m^{t,i} \cdot \nabla F_m(\boldsymbol{w}_m^{t,i-1}; \xi_m^{t,i})\right\|^2\right\} \tag{41b}$$

$$\leqslant L^2 \sum_{k=2}^{K^t} \left(\sum_{i=1}^{k-1} (\eta_m^{t,i})^2 \cdot \sum_{i=1}^{k-1} \mathbb{E}\left\{\left\|\nabla F_m(\boldsymbol{w}_m^{t,i-1}; \xi_m^{t,i})\right\|^2\right\}\right) \tag{41c}$$

$$\leqslant L^2 \sum_{k=2}^{K^t} \left(\sum_{i=1}^{k-1} (\eta_m^{t,i})^2 \cdot (k-1)(G^2 + \sigma^2)\right) \tag{41d}$$

$$= L^2(G^2 + \sigma^2) \sum_{k=2}^{K^t} (k-1) \sum_{i=1}^{k-1} (\eta_m^{t,i})^2 \tag{41e}$$

where

- the inequality (41a) holds according to Assumption 1;
- the inequality (41c) comes from Cauchy-Schwarz inequality $(a_1 b_1 + a_2 b_2 + a_3 b_3 + \cdots + a_n b_n)^2 \leqslant (a_1^2 + a_2^2 + a_3^2 + \cdots + a_n^2)(b_1^2 + b_2^2 + b_3^2 + \cdots + b_n^2)$;
- the inequality (41d) can be derived from Assumptions 3, 4 and 5.

Combining (20), (38), (40) and (41), we obtain the following inequality

$$\mathbb{E}\left\{\left\|\boldsymbol{z}^t - \nabla F(\boldsymbol{w}^t)\right\|^2\right\} \leqslant \Delta^t + \frac{8\sigma^2 (C_\alpha)^2}{(2C_\alpha - 1)^2} + \frac{2\epsilon^2}{(2C_\alpha - 1)^2} \tag{42}$$

This completes the proof of Lemma 1.

Appendix D
Proof of Lemma 2

With Lemma 3, there is

$$\mathbb{E}\left\{\left\|\boldsymbol{z}^t\right\|^2\right\} \leqslant \frac{8C_\alpha}{(2C_\alpha - 1)^2} \sum_{m \in \mathcal{M} \setminus \mathcal{B}} \alpha_m \mathbb{E}\left\|\boldsymbol{z}_m^t\right\|^2 + \frac{2\epsilon^2}{(2C_\alpha - 1)^2} \tag{43}$$

For the term $\mathbb{E}\left\{\left\|\boldsymbol{z}_m^t\right\|^2\right\}$, $m \in \mathcal{M} \setminus \mathcal{B}$, according to Assumptions 3, 4 and 5, we have

$$\mathbb{E}\left\{\left\|\boldsymbol{z}_m^t\right\|^2\right\} = \mathbb{E}\left\{\left\|\frac{1}{K^t}\sum_{k=1}^{K^t}\nabla F_m(\boldsymbol{w}_m^{t,k-1};\xi_m^{t,k})\right\|^2\right\} \tag{44a}$$

$$\leqslant \frac{1}{K^t}\sum_{k=1}^{K^t}\mathbb{E}\left\{\left\|\nabla F_m(\boldsymbol{w}_m^{t,k-1};\xi_m^{t,k})\right\|^2\right\} \tag{44b}$$

$$\leqslant G^2 + \sigma^2 \tag{44c}$$

where the inequality (44b) comes from Cauchy-Schwarz inequality $(a_1 + a_2 + a_3 + \cdots + a_n)^2 \leqslant n(a_1^2 + a_2^2 + a_3^2 + \cdots + a_n^2)$. Hence we can get the following inequality

$$\mathbb{E}\left\{\left\|\boldsymbol{z}^t\right\|^2\right\} \leqslant \frac{8(C_\alpha)^2(G^2+\sigma^2)}{(2C_\alpha - 1)^2} + \frac{2\epsilon^2}{(2C_\alpha - 1)^2} \tag{45}$$

Since $\mathbb{E}\left\{\left\|\boldsymbol{z}^t\right\|^2\right\} \geqslant 0$, then there is

$$0 \leqslant \mathbb{E}\left\{\left\|\boldsymbol{z}^t\right\|^2\right\} \leqslant \frac{8(C_\alpha)^2(G^2+\sigma^2)}{(2C_\alpha - 1)^2} + \frac{2\epsilon^2}{(2C_\alpha - 1)^2} \tag{46}$$

This completes the proof of Lemma 2.

APPENDIX E
PROOF OF LEMMA 3

For the ease of presentation, we denote

$\boldsymbol{z}^* = \text{geomed}\left(\{\boldsymbol{z}_m | m \in \mathcal{M}\}\right)$ and

$\boldsymbol{z}_\epsilon^* = \text{geomed}\left(\{\boldsymbol{z}_m | m \in \mathcal{M}\}, \epsilon\right)$, then there is

$$\sum_{i=1}^{M}\frac{S_i}{\sum_{j=1}^{M}S_j}\left\|\boldsymbol{z}_\epsilon^* - \boldsymbol{z}_i\right\| \leqslant \sum_{i=1}^{M}\frac{S_i}{\sum_{j=1}^{M}S_j}\left\|\boldsymbol{z}^* - \boldsymbol{z}_i\right\| + \epsilon \tag{47}$$

which implies

$$\sum_{i=1}^{M}\alpha_i\left\|\boldsymbol{z}_\epsilon^* - \boldsymbol{z}_i\right\| \leqslant \sum_{i=1}^{M}\alpha_i\left\|\boldsymbol{z}^* - \boldsymbol{z}_i\right\| + \epsilon \tag{48}$$

by recalling the definition of α_i as given in (17).

By the definition of geometric median, there is

$$\sum_{i=1}^{M}\alpha_i\left\|\boldsymbol{z}^* - \boldsymbol{z}_i\right\| = \inf_{\boldsymbol{y}}\sum_{i=1}^{M}\alpha_i\left\|\boldsymbol{y} - \boldsymbol{z}_i\right\| \leqslant \sum_{i=1}^{M}\alpha_i\left\|\boldsymbol{z}_i\right\| \tag{49}$$

According to the triangle inequality, we have

$$\sum_{\boldsymbol{z}_i \in \mathcal{Z}'} \alpha_i \left\| \boldsymbol{z}_\epsilon^* - \boldsymbol{z}_i \right\| \geqslant \sum_{\boldsymbol{z}_i \in \mathcal{Z}'} \alpha_i \left(\left\| \boldsymbol{z}_i \right\| - \left\| \boldsymbol{z}_\epsilon^* \right\| \right) \tag{50}$$

$$\sum_{\boldsymbol{z}_i \in \mathcal{Z} \setminus \mathcal{Z}'} \alpha_i \left\| \boldsymbol{z}_\epsilon^* - \boldsymbol{z}_i \right\| \geqslant \sum_{\boldsymbol{z}_i \in \mathcal{Z} \setminus \mathcal{Z}'} \alpha_i \left(\left\| \boldsymbol{z}_\epsilon^* \right\| - \left\| \boldsymbol{z}_i \right\| \right) \tag{51}$$

By adding (50) and (51) together, we obtain

$$\begin{aligned} \sum_{\boldsymbol{z}_i \in \mathcal{Z}} \alpha_i \left\| \boldsymbol{z}_\epsilon^* - \boldsymbol{z}_i \right\| \geqslant \sum_{\boldsymbol{z}_i \in \mathcal{Z}} \alpha_i \left\| \boldsymbol{z}_i \right\| + (1 - &\sum_{\boldsymbol{z}_i \in \mathcal{Z}'} 2\alpha_i) \left\| \boldsymbol{z}_\epsilon^* \right\| \\ &- 2 \sum_{\boldsymbol{z}_i \in \mathcal{Z} \setminus \mathcal{Z}'} \alpha_i \left\| \boldsymbol{z}_i \right\| \end{aligned} \tag{52}$$

Combining (48), (49), and (52), and with the fact that $\sum_{\boldsymbol{z}_i \in \mathcal{Z}'} \alpha_i = 1 - C_\alpha < 1/2$, we have the following inequality

$$\left\| \boldsymbol{z}_\epsilon^* \right\| \leqslant \frac{2 \sum_{\boldsymbol{z}_i \in \mathcal{Z} \setminus \mathcal{Z}'} \alpha_i \left\| \boldsymbol{z}_i \right\| + \epsilon}{1 - 2 \sum_{\boldsymbol{z}_i \in \mathcal{Z}'} \alpha_i} \tag{53}$$

Squaring both sides of (53) and based on the Cauchy-Schwarz inequality, we further have

$$\begin{aligned} &\left\| \boldsymbol{z}_\epsilon^* \right\|^2 \\ &\leqslant \left(\frac{2 \sum_{\boldsymbol{z}_i \in \mathcal{Z} \setminus \mathcal{Z}'} \alpha_i \left\| \boldsymbol{z}_i \right\| + \epsilon}{1 - 2 \sum_{\boldsymbol{z}_i \in \mathcal{Z}'} \alpha_i} \right)^2 &\text{(54a)} \\ &\leqslant \frac{8 \left(\sum_{\boldsymbol{z}_i \in \mathcal{Z} \setminus \mathcal{Z}'} \alpha_i \left\| \boldsymbol{z}_i \right\| \right)^2 + 2\epsilon^2}{\left(1 - 2 \sum_{\boldsymbol{z}_i \in \mathcal{Z}'} \alpha_i \right)^2} &\text{(54b)} \\ &= \frac{8 \left(\sum_{\boldsymbol{z}_i \in \mathcal{Z} \setminus \mathcal{Z}'} \alpha_i \right)^2}{\left(1 - 2 \sum_{\boldsymbol{z}_i \in \mathcal{Z}'} \alpha_i \right)^2} \left(\sum_{\boldsymbol{z}_i \in \mathcal{Z} \setminus \mathcal{Z}'} \frac{\alpha_i}{\sum_{\boldsymbol{z}_j \in \mathcal{Z} \setminus \mathcal{Z}'} \alpha_j} \left\| \boldsymbol{z}_i \right\| \right)^2 + \frac{2\epsilon^2}{\left(1 - 2 \sum_{\boldsymbol{z}_i \in \mathcal{Z}'} \alpha_i \right)^2} &\text{(54c)} \\ &\leqslant \frac{8 \left(\sum_{\boldsymbol{z}_i \in \mathcal{Z} \setminus \mathcal{Z}'} \alpha_i \right) \sum_{\boldsymbol{z}_i \in \mathcal{Z} \setminus \mathcal{Z}'} \alpha_i \left\| \boldsymbol{z}_i \right\|^2 + 2\epsilon^2}{\left(1 - 2 \sum_{\boldsymbol{z}_i \in \mathcal{Z}'} \alpha_i \right)^2} &\text{(54d)} \end{aligned}$$

where

- the inequality (54b) comes from $(a+b)^2 \leqslant 2a^2 + 2b^2$;
- the inequality (54d) holds because of Jensen's inequality $f\left(\sum_{i=1}^{n} \nu_i x_i \right) \leqslant \sum_{i=1}^{n} \nu_i f(x_i)$ with $\sum_{i=1}^{n} \nu_i = 1$, $\nu_i \geq 0$ for $i = 1, 2, ..., n$, and $f(x) = x^2$ being a convex function.

Recalling of the definition of C_α in (17), we then have

$$\left\|\boldsymbol{z}_\epsilon^*\right\|^2 \leqslant \frac{8C_\alpha}{\left(2C_\alpha-1\right)^2}\sum_{\boldsymbol{z}_i\in\mathcal{Z}\setminus\mathcal{Z}'}\alpha_i\left\|\boldsymbol{z}_i\right\|^2+\frac{2\epsilon^2}{\left(2C_\alpha-1\right)^2} \tag{55}$$

Taking the expectation of the both sides of (55) can lead to the statement of Lemma 3.

References

[1] L. Zhou, K.-H. Yeh, G. Hancke, Z. Liu, and C. Su, "Security and privacy for the industrial internet of things: An overview of approaches to safeguarding endpoints," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 76–87, 2018.

[2] H. B. McMahan, F. Yu, P. Richtarik, A. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *Proc. Adv. Conf. Neural Inf. Process. Syst.*, pp. 5–10, 2016.

[3] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proc. Mach. Learn. Syst.*, pp. 429–450, 2020.

[4] R. Agrawal and R. Srikant, "Privacy-preserving data mining," *Proc. ACM SIGMOD Int. Conf. Manage. Data*, pp. 439–450, 2000.

[5] J. Duchi, M. J. Wainwright, and M. I. Jordan, "Local privacy and minimax bounds: Sharp rates for probability estimation," *Proc. Int. Conf. Neural Inf. Process. Syst.*, pp. 1529–1537, 2013.

[6] J. Konečnỳ, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv preprint arXiv:1610.02527*, 2016.

[7] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He, "A survey on federated learning systems: Vision, hype and reality for data privacy and protection," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 3347 – 3366, 2021.

[8] 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.*, vol. 37, no. 6, pp. 1205–1221, 2019.

[9] M. Chen, D. Gündüz, K. Huang, W. Saad, M. Bennis, A. V. Feljan, and H. V. Poor, "Distributed learning in wireless networks: Recent progress and future challenges," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3579–3605, 2021.

[10] A. Vempaty, L. Tong, and P. K. Varshney, "Distributed inference with byzantine data: State-of-the-art review on data falsification attacks," *IEEE Signal Process. Mag.*, vol. 30, no. 5, pp. 65–75, 2013.

[11] Z. Yang, A. Gang, and W. U. Bajwa, "Adversary-resilient distributed and decentralized statistical inference and machine learning: An overview of recent advances under the byzantine threat model," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 146–159, 2020.

[12] J. So, B. Güler, and A. S. Avestimehr, "Byzantine-resilient secure federated learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2168–2181, 2020.

[13] X. Cao and L. Lai, "Distributed gradient descent algorithm robust to an arbitrary number of byzantine attackers," *IEEE Trans. Signal Process.*, vol. 67, no. 22, pp. 5850–5864, 2019.

[14] Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," *Proc. ACM Meas. Anal. Comput. Syst.*, pp. 1–25, 2017.

[15] X. Cao and L. Lai, "Distributed approximate newton's method robust to byzantine attackers," *IEEE Trans. Signal Process.*, vol. 68, pp. 6011–6025, 2020.

[16] C. Xie, O. Koyejo, and I. Gupta, "Generalized byzantine-tolerant SGD," *arXiv preprint arXiv:1802.10116*, 2018.

[17] 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.

[18] Z. Xie and S. Song, "FedKL: Tackling data heterogeneity in federated reinforcement learning by penalizing kl divergence," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 4, pp. 1227–1242, 2023.

[19] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," *Proc. Int. Conf. Machine Learning*, pp. 5650–5659, 2018.

[20] L. Su and J. Xu, "Securing distributed gradient descent in high dimensional statistical learning," *Proc. ACM Meas. Anal. Comput. Syst.*, no. 1, pp. 1–41, 2019.

[21] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, pp. 118–128, 2017.

[22] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "BEV-SGD: Best effort voting SGD against byzantine attacks for analog-aggregation-based federated learning over the air," *IEEE Internet Things J.*, vol. 9, no. 19, pp. 18946–18959, 2022.

[23] L. Li, W. Xu, T. Chen, G. B. Giannakis, and Q. Ling, "RSA: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets," *Proc. AAAI Conf. Artif. Intell.*, no. 01, pp. 1544–1551, 2019.

[24] B. Turan, C. A. Uribe, H.-T. Wai, and M. Alizadeh, "Robust distributed optimization with randomly corrupted gradients," *IEEE Trans. Signal Process.*, vol. 70, pp. 3484–3498, 2022.

[25] K. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust aggregation for federated learning," *IEEE Trans. Signal Process.*, vol. 70, pp. 1142–1154, 2022.

[26] Z. Wu, Q. Ling, T. Chen, and G. B. Giannakis, "Federated variance-reduced stochastic gradient descent with robustness to byzantine attacks," *IEEE Trans. Signal Process.*, vol. 68, pp. 4583–4596, 2020.

[27] H. Zhu and Q. Ling, "Byzantine-robust distributed learning with compression," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 9, pp. 280 – 294, 2023.

[28] S. Minsker, "Geometric median and robust estimation in banach spaces," *Bernoulli*, pp. 2308–2335, 2015.

[29] A. Defazio, F. Bach, and S. Lacoste-Julien, "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives," *Proc. Adv. Neural Inf. Process. Syst.*, pp. 1646–1654, 2014.

[30] M. Huang, D. Zhang, and K. Ji, "Achieving linear speedup in non-iid federated bilevel learning," *arXiv preprint arXiv:2302.05412*, 2023.

[31] X. Li and P. Li, "Analysis of error feedback in federated non-convex optimization with biased compression," *arXiv preprint arXiv:2211.14292*, 2022.

[32] K. Aftab, R. Hartley, and J. Trumpf, "Generalized weiszfeld algorithms for lq optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 728–745, 2014.

[33] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[34] K. Yue, R. Jin, R. Pilgrim, C.-W. Wong, D. Baron, and H. Dai, "Neural tangent kernel empowered federated learning," *Proc. Int. Conf. Mach. Learn.*, pp. 25783–25803, 2022.