Abstract—In this article, we study the problem of Byzantine fault-tolerance in a federated optimization setting, where there is a group of agents communicating with a centralized coordinator. We allow up to $f$ Byzantine-faulty agents, which may not follow a prescribed algorithm correctly, and may share arbitrarily incorrect information with the coordinator. Associated with each nonfaulty agent is a local cost function. The goal of the nonfaulty agents is to compute a minimizer of their aggregate cost function. For solving this problem, we propose a local gradient-descent algorithm that incorporates a novel comparative elimination filter (aka. aggregation scheme) to provably mitigate the detrimental impact of Byzantine faults. In the deterministic setting, when the agents can compute their local gradients accurately, our algorithm guarantees exact fault-tolerance against a bounded fraction of Byzantine agents, provided the nonfaulty agents satisfy the known necessary condition of $2f$-redundancy. In the stochastic setting, when the agents can only compute stochastic estimates of their gradients, our algorithm guarantees approximate fault-tolerance where the approximation error is proportional to the variance of stochastic gradients and the fraction of Byzantine agents.

Index Terms—$2f$-redundancy, Byzantine fault-tolerance, comparative elimination (CE), federated learning.

I. INTRODUCTION

We consider a distributed optimization framework where there are $N$ agents communicating with a single coordinator. This framework is also popularly known as federated optimization [1]. Associated with each agent $i$ is a function $q^i: \mathbb{R}^d \to \mathbb{R}$. The goal of the agents is to find $x^*$ such that

$$x^* = \arg \min_{x \in \mathbb{R}^d} \sum_{i=1}^{N} q^i(x)$$

(1)

Manuscript received 24 June 2022; revised 7 November 2022; accepted 27 December 2022. Date of publication 16 January 2023; date of current version 5 December 2023. This work was supported in part by ARL through Cooperative Agreement under Grant W911NF-17-2-0196. The work of Thinh T. Doan was supported in part by Commonwealth Cyber Initiative, an investment in the advancement of cyber R&D, innovation, and workforce development. Recommended by Associate Editor K. You. (Corresponding author: Nirupam Gupta.)

Nirupam Gupta is with the School of Computer and Communication Sciences, EPFL, 1015 Lausanne, Switzerland (e-mail: nirupam.gupta@epfl.ch).

Thinh T. Doan is with the Bradley Department of Electrical and Computer Department, Virginia Tech, Blacksburg, VA 24060 USA (e-mail: thinhdt210@gmail.com).

Nitin Vaidya is with the Department of Computer Science, Georgetown University, Washington, DC 20057-1232 USA (e-mail: nv198@georgetown.edu).

Digital Object Identifier 10.1109/TCNS.2023.3237489

where each $q^i$ is given as

$$q^i(x) = \mathbb{E}_{\pi_i}[Q^i(x; X^i)]$$

(2)

for some random variable $X^i$ defined over a sample set $\mathcal{X}$ with a distribution $\pi^i$. We assume that each agent $i$ only has access to a sequence of vectors $\{G^i(\cdot)\}$, which can be either the actual gradients $\{\nabla q^i(\cdot)\}$ or stochastic gradients $\{\nabla Q^i(\cdot; X^i)\}$ of its cost function $q^i(\cdot)$. This is a common distributed setting for inferring from data distributed across different agents. The goal is to design an algorithm that allows these agents to jointly minimize a loss function defined over their data, i.e., solve the optimization problem (1). Besides traditional machine-learning applications [2], federated optimization now also finds application in networked systems, e.g., Internet of Vehicles [3] and industrial control systems [4], [5]. For solving this problem, we consider the federated local stochastic gradient-descent (local SGD) method, which has recently received significant attention due to its application in distributed learning [2], [6]. In this method, the coordinator maintains an estimate of a solution $x^*$ defined in (1). This estimate is broadcast to all the agents, and each agent updates its copy of the estimate by running a number of local SGD steps. The agents send back to the coordinator their updated estimates. Finally, the coordinator averages the received estimates to obtain a new global estimate of $x^*$. Eventually, if all the agents are nonfaulty, the sequence of global estimates converges to a solution (1). Since the agents only share their local estimates and not their data, federated local SGD is widely used in privacy conscious distributed learning [6].

Our interest is to study the federated local SGD algorithm when up to $f$ agents are Byzantine faulty [7]. Such faulty agents may behave arbitrarily, and their identity is a priori unknown. In particular, the Byzantine faulty agents may share incorrect information with the coordinator in order to corrupt the output of the algorithm, e.g., see [8]. The Byzantine fault model covers both the computational and communication errors that may occur due to software and hardware bugs or an adversary controlling part of the system [9]. We aim to design a new local SGD method that allows all the nonfaulty agents to compute an exact minimum of the aggregate cost of the nonfaulty agents, despite the presence of the Byzantine agents. Specifically, we consider the problem of exact fault-tolerance defined as follows. For a set $\mathcal{H}$, $|\mathcal{H}|$ denotes its cardinality.
**Definition 1:** Let $\mathcal{H}$ be the set of nonfaulty agents. A distributed optimization algorithm is said to have exact fault-tolerance if it allows all the nonfaulty agents to compute

$$x^*_H \in \arg \min_{x \in \mathbb{R}^d} \sum_{i \in \mathcal{H}} q^i(x).$$  \hspace{1cm} (3)

Since the identity of the Byzantine agents is a priori unknown, in general, exact fault-tolerance is unachievable [10]. Indeed, recent work has shown that exact fault-tolerance is achievable if and only if the nonfaulty agents satisfy the property of $2f$-redundancy defined as follows (proposed in [11]).

**Definition 2:** A set of nonfaulty agents $\mathcal{H}$ is said to have 2f-redundancy if for any subset $S \subseteq \mathcal{H}$ with $|S| \geq N-2f$

$$\arg \min_{x \in \mathbb{R}^d} \sum_{i \in S} q^i(x) = \arg \min_{x \in \mathbb{R}^d} \sum_{i \in \mathcal{H}} q^i(x).$$  \hspace{1cm} (4)

The 2f-redundancy property is critical to our algorithm, presented in Section II, for solving (3). Formal proof for the necessity of 2f-redundancy can be found in [11]. Essentially, this property implies that a minimizer of the aggregate cost of any $N-2f$ nonfaulty agents is also a minimizer of the aggregate cost of all the nonfaulty agents, and vice-versa. This seemingly contrived condition arises naturally with high probability in many practical applications of the distributed optimization problem, such as distributed hypothesis testing [12], [13], [14], [15] and distributed learning [16], [17], [18], [19].

In the context of distributed learning, when all agents have the same data generating distribution (i.e., the homogeneous data setting), 2f-redundancy holds true trivially. In general, however, 2f-redundancy also holds true in the heterogeneous data setting if we can solve the nonfaulty distributed learning problem (3) using only $N-2f$ nonfaulty agents.

Prior work indicates that solving the exact fault-tolerance problem (3) is nontrivial even under the 2f-redundancy property, especially in the high dimension case, e.g., see [11], [15], and [20]. This can be attributed to two main factors: 1) the identity of the Byzantine faulty agents is a priori unknown, and 2) the smart Byzantine agents can inject malicious information without getting detected (e.g., see [21]). Although there exist techniques that impart provable exact fault-tolerance to the distributed implementation of SGD algorithm (in which the agents simply share their local gradients, and do not maintain local estimates) [11], [17], [22], we do not know of such techniques for the federated local SGD algorithm. This motivates us to propose a technique named comparative elimination (CE) that provably robustifies the federated local SGD algorithm against the Byzantine agents. In the special deterministic setting, we show that CE can provide exact fault-tolerance. In the generic stochastic setting, we achieve approximate fault-tolerance with approximation error proportional to the variance of stochastic gradients and fraction of Byzantine agents $f/N$.

Intuitively speaking, the CE technique allows the coordinator to mitigate the detrimental impact of possibly adversarial estimates sent by the Byzantine agents. Specifically, instead of simply averaging the agents’ local estimates, the coordinator eliminates $f$ estimates that are farthest from the current global estimate. Then, the remaining $N-f$ agents’ estimates are averaged to obtain the new global estimate. Our scheme and its fault-tolerance properties are formally presented in Section II.

**A. Main Contributions**

We formally analyze the Byzantine robustness of the federated local SGD algorithm when coupled with the aforementioned technique of CE. Specifically, assuming each nonfaulty agent’s cost function to be $L$-smooth, the aggregate nonfaulty cost to be $\mu$-strongly convex (but, the local costs may only be convex), and the necessary condition of 2f-redundancy, we present the following results.

1) In the deterministic case, when each nonfaulty agent $i$ updates its local estimates using actual gradients of its cost $\{\nabla q^i(\cdot)\}$, the CE filter scheme guarantees exact fault-tolerance if $\frac{f}{|\mathcal{H}|} \leq \frac{\mu}{2L}$. Moreover, the convergence is linear, similar to the fault-free setting under strong convexity.

2) In the stochastic case, when a nonfaulty agent $i$ can only compute unbiased noisy estimates of its local gradients $\{\nabla Q^i(\cdot, X^i)\}$ we guarantee approximate fault-tolerance. Specifically, the sequence of global estimates $\{\tilde{x}_k\}$ satisfy the following for all $k$:

$$E [\|\tilde{x}_k - x^*_H\|^2] \leq \lambda^k E [\|\bar{x}_0 - x^*_H\|^2] + O \left(\frac{\sigma^2 \alpha + \sigma^2 f}{N - f}\right)$$

for some $\lambda \in (0, 1)$ where $\alpha$ is a constant step-size of the algorithm, and $\sigma$ is the variance of stochastic gradients.

Specific details of our results are given in Section III. **Other applications of comparative elimination:** The process of eliminating or clipping the outlying updates in any iteration of a gradient-based optimization method, similar to that in CE, has also been shown useful in addressing other issues. For example, the authors in [23] and [24] use a similar approach for approximately solving the NP-hard problem of quadratic system of equations. Threshold clipping has also been used for mitigating the problem of gradient explosion in training of neural networks [25], and differential privacy preservation [26]. Nevertheless, the applicability of this approach to the Byzantine robustness of federated local SGD is novel, and analyzing its correctness remains nontrivial.

**B. Other Related Work**

In recent years, several schemes have been proposed for the Byzantine fault-tolerance in distributed implementation of SGD algorithm. Prominent schemes include coordinatewise trimmed mean (CWTM) [10], [15], [22], [27], multi-KRM [17], geometric median of means (GMoM) [28], coordinatewise median [22], [29], Bulyan [19], minimum-diameter averaging (MDA) [19], phocas [30], Byzantine-robust stochastic aggregation (RSA) [31], signSGD with majority voting [32], and spectral decomposition-based filters [33], [34]. Most of these works, with the exception of [10], [15], [20], [27], [29], and [31], only consider the framework of distributed SGD, wherein the agents send gradients of their local costs, and their results are not readily applicable to the federated local SGD framework that
we consider. Nevertheless, these works suggest that the aforementioned schemes need not guarantee exact fault-tolerance in general, even in the deterministic setting with 2f-redundancy, unless further assumptions are made on the nonfaulty agents’ costs.

Although [9], [10], [29] implicitly show exact fault-tolerance properties of trimmed-mean and median, they only consider the scalar case, where the agents’ cost functions are univariate. The extension of their results to higher-dimensions is nontrivial and remains poorly understood, e.g., see [15], [20], [27], [35]. For instance, Su and Vaidya [35] considered a degenerate case, wherein the agents’ cost functions have a known common basis. The prior work [15] shows that CWTM can guarantee exact fault-tolerance when solving the distributed linear least squares problem provided the agents’ data satisfies a condition stronger than 2f-redundancy. In [27], they assume that the agents’ costs can be decomposed into independent scalar strictly convex functions. Recently, [20] studied the fault-tolerance of CWTM in a peer-to-peer setting (a generalization of federated model) for generic convex optimization problems; their results suggest that CWTM need not provide exact fault-tolerance even under 2f-redundancy.

When applied to the federated local SGD framework, some of the above schemes, including multi-KRUM, Bulyan, CWTM, GMoM, and MDA, operate only on the local estimates sent by the agents and disregard the current global estimate maintained by the coordinator [36]. On the other hand, our proposed CE filter exploits the (supposed) closeness between the current global estimate and the nonfaulty agents’ local updated estimates to improve robustness against the Byzantine agents. The closeness between the global and nonfaulty agents’ local estimates exists due to the Lipschitz smoothness of agents’ local cost functions. This is a critical observation for the fault-tolerance property of our algorithm. Other works that also exploit this observation are RSA [31], [37].

Recently, it has shown that geometric median aggregation scheme provably provides improved Byzantine fault-tolerance compared to other aggregation schemes in federated model [38]. However, computing geometric median is a challenging problem, as there does not exist a closed-form formula [39]. Moreover, existing numerical algorithms for computing geometric median are only approximate, and computationally quite complex [40]. Other schemes, such as the verifiable coding in [41] and manual verification of information sent by agents [42], are not directly applicable to the commonly used federated framework where interagent communication is absent or there are a large number of agents, and data privacy is a major concern.

Besides federated local SGD framework, the proposed CE filter can also guarantee exact fault-tolerance in the distributed SGD framework, where agents share their gradients instead of estimates [11], [43]. Also, for the distributed SGD framework, recent works have shown that momentum helps improve the fault-tolerance of a Byzantine-robust aggregation scheme [44], [45], [46]. However, adapting these results to federated local SGD is nontrivial and remains to be investigated.

Additional results characterizing the impact of redundancy on Byzantine fault-tolerance in distributed optimization can be found in [11], [43], and [47]. An extension of the 2f-redundancy property has also been shown vital in analyzing the solvability of the optimization problem (3) when the nonfaulty agents coordinate asynchronously [48]. However, these works consider the problem of the Byzantine robustness in distributed SGD, where the agents share gradients instead of locally updated parameters, and their proposed algorithms cannot be trivially extended to the federated local SGD setting.

II. LOCAL SGD UNDER BYZANTINE MODEL

We now present the proposed algorithm for solving (1) in the presence of at most f Byzantine faulty agents. We note that the Byzantine agents can observe the values of other agents and send arbitrarily values to the coordinator. To handle this scenario, the main idea of our approach is a Byzantine robust aggregation rule (or filter), named comparative elimination (CE) filter, which is implemented at the coordinator. This filter together with the local SGD formulates our proposed method, formally presented in Algorithm 1 for solving (3).

In Algorithm 1, each agent i maintains a local variable x^i, and the coordinator maintains ̄x, the average of these x^i. At any iteration k ≥ 0, agent i receives x_k from the coordinator and initializes its iterate x^i_k,0 = x_k. Here x^i_k denotes the iterate at iteration k and local time t ∈ [0,...,T − 1] at agent i. Agent i then runs a number T of local SGD steps using time-varying step sizes α_k and its local direction ∇^i_k(x^i), which can be either the actual gradient ∇q^i(x^i_k,t) or a stochastic estimate ∇Q^i(·,X^i) of its gradient based on the data {X^i_k,t} sampled i.i.d from π^i. After T local SGD steps, the agents then send their new local updates x^i,T to the coordinator. However, Byzantine agents may send arbitrary values to disrupt the learning process. The coordinator implements the CE filter [in steps 2(a) and 2(b) of Algorithm 1] to dilute the impact of “bad” values sent by the Byzantine agents. The main idea of this filter is to discard f-values (or estimates) that are f-farthest from the current global estimate ̄x_k. Finally, the coordinator averages the N − f remaining estimates, as shown in (7), to compute the new global estimate. Note that without the CE filter [i.e., without steps 2(a) and 2(b), and F^i_k = [1,N]], Algorithm 1 reduces to the traditional local GD method.

III. MAIN RESULTS

In this section, we present the main results of this article, where we characterize the convergence of Algorithm 1 for solving problem (3). We consider two cases, namely, the deterministic settings (when G^i(·) = ∇q^i(·)) and the stochastic settings (when G^i(·) = ∇Q^i(·,X^i))1. In both cases, our theoretical results are derived when the nonfaulty agents’ cost functions are smooth. Moreover, we also assume the average nonfaulty cost

1Proofs of all the theorems presented in this section are deferred to the Appendix attached after the list of references.
Algorithm 1: Local SGD With CE Filter.

**Initialization:** The coordinator initializes $\bar{x}_0 \in \mathbb{R}^d$. Agent $i$ initializes step sizes $\{\alpha_k\}$ and a positive integer $T$.

**Iterations:** For $k = 0, 1, 2, \ldots$

1) Agent $i$
   a) Receive $\bar{x}_k$ sent by the server and set $x_{k,0} = \bar{x}_k$
   b) For $t = 0, 1, \ldots, T - 1$, implement
      \[
      x_{k,t+1}^i = x_{k,t}^i - \alpha_k G^i(x_{k,t}^i).
      \]  
2) The coordinator receives $x_{k,T}^i$ from each agent $i$ and implements the CE filter as follows.
   a) Compute the distances of $x_{k,T}^i$ from its current value $\bar{x}_k$, and sort them in an increasing order
      \[
      \|\bar{x}_k - x_{k,T}^1\| \leq \ldots \leq \|\bar{x}_k - x_{k,T}^{|\mathcal{H}|}\|.
      \]  
   b) Discard the $j$-largest distances, i.e., it drops $x_{k,T}^{L-1}, \ldots, x_{k,T}^{|\mathcal{H}|}$. Let $F_k = \{i_1, \ldots, i_{N-j}\}$.
   c) Update its iterate as
      \[
      \bar{x}_{k+1} = \frac{1}{|F_k|} \sum_{i \in F_k} x_{k,T}^i.
      \]  

function, denoted by $q^H(x)$, to be strongly convex. Specifically,
\[
q^H(x) = \frac{1}{|H|} \sum_{i \in H} q^i(x).
\]  

These assumptions are formally stated as follows.

**Assumption 1. (Lipschitz smoothness):** The nonfaulty agents’ functions have Lipschitz continuous gradients, i.e., there exists a positive constant $L < \infty$ such that $\forall i \in H$
\[
\|\nabla q^i(x) - \nabla q^i(y)\| \leq L\|x - y\| \quad \forall x, y \in \mathbb{R}^d.
\]  

**Assumption 2. (Strong convexity):** $q^H$ is strongly convex, i.e., there exists a positive constant $\mu < \infty$ such that
\[
(x - y)^T (\nabla q^H(x) - \nabla q^H(y)) \geq \mu \|x - y\|^2 \quad \forall x, y \in \mathbb{R}^d
\]  

where $(\cdot)^T$ denotes the transpose.

To this end, we assume that these assumptions and the $2f$-redundancy property always hold true. In addition, without loss of generality we consider $|H| = N - f$. Finally, note that Assumptions 1 and 2 hold true simultaneously only if $\mu \leq L$.

**Remark 1:** Assumption 2 implies that there exists a unique solution $x^*_H$ of problem (3). However, this assumption does not imply that each local function $q^i$ is strongly convex. Indeed, each $q^i$ can have more than one minimizer. Under the $2f$-redundancy property one can show that
\[
x^*_H \in \bigcap_{i \in H} \arg \min_{x \in \mathbb{R}^d} q^i(x).
\]  

Thus, one can view that Algorithm 1 tries to seek a point in the intersection of the minimizer sets of the local functions $q^i$. However, we do not assume we can compute these sets since the task is intractable in general. Our analysis given later relies on (9), whose proof can be found in [11, Appendix B].

### A. Deterministic Settings

In this section, we consider the deterministic setting of Algorithm 1, i.e., $G^i(\cdot) = \nabla q^i(\cdot)$. To present the main idea of our analysis in handling the Byzantine agents, we first study the convergence of Algorithm 1 when $T = 1$ in Section III-A1 and generalize to the case $T > 1$ in Section III-A2.

1) **Case of $T = 1$:** When $T = 1$, Algorithm 1 is equivalent to the popular distributed (stochastic) gradient method. Indeed, by (5) we have for any $i \in \mathcal{H}$
\[
x_{k,1}^i = x_{k,0}^i - \alpha_k \nabla q^i(x_{k,0}^i) = \bar{x}_k - \alpha_k \nabla q^i(\bar{x}_k).
\]  

We denote by $B$ the set of Byzantine agents, i.e., $N = |B| + |\mathcal{H}|$ and $|B| \leq f$. Without loss of generality we assume that $|B| = f$. Similarly, let $B_k$ be the set of Byzantine agents in $F_k$ and $\mathcal{H}_k$ be the set of nonfaulty agents in $F_k$. Then we have $|B_k| = |F_k \setminus \mathcal{H}_k| \leq f$, for any $k \geq 0$.

**Theorem 1:** Let $\{\bar{x}_k\}$ be generated by Algorithm 1 with $T = 1$. We assume that the following condition holds:
\[
\frac{f}{N - f} \leq \frac{\mu}{3L}.
\]  

Let $\alpha_k$ be chosen as
\[
\alpha_k = \alpha \leq \frac{\mu}{4L^2}.
\]  

Then we have
\[
\|\bar{x}_k - x^*_H\|^2 \leq \left(1 - \frac{\mu\alpha}{6}\right)^k \|\bar{x}_0 - x^*_H\|^2.
\]  

**Proof:** To prove the theorem, we begin by analyzing the growth of the squared distance between $\bar{x}_k$ and $x^*_H$, i.e., $\|\bar{x}_k - x^*_H\|^2$, along the trajectory of Algorithm 1. We first express it in terms of the squared distance between the local estimates of agents that were not filtered-out by CE, i.e., $\|x_{k,1}^i - x^*_H\|^2$, where $i \in F_k$. However, some of the agents in $F_k$ might actually be Byzantine. This is where we use the fact that CE filter eliminates $f$ estimates that are farthest from $\bar{x}_k$. That is, for each Byzantine agent $j \in F_j$, there exists a nonfaulty agent $i \in \{1, \ldots, N\} \setminus F_k$ such that $\|x_{k,1}^i - \bar{x}_k\| \leq \|x_{k,1}^j - \bar{x}_k\|$. This observation allows us to bound the impact of any incorrect estimates sent by the Byzantine agents. Finally, we combine this observation with the fact that the nonfaulty agents have a common stationary point under $2f$-redundancy (see [6]). But, as each nonfaulty agent may still have nonunique stationary points, to ultimately ensure convergence of the average nonfaulty gradients to stationarity we make use of the assumption on strong convexity of the average nonfaulty cost function.

The full proof of Theorem 1 is deferred to Appendix A. The main ideas behind the proofs for the subsequent results presented in Theorem 2 and Theorem 3 remain similar to that of Theorem 1 presented above.

**Remark 2:** In Theorem 1 we show that under the $2f$ redundancy, Algorithm 1 returns an exact solution $x^*_H$ of problem (3) even under of at most $f$ Byzantine agents. Moreover, the convergence is linear, which is the same as what we expect in the nonfaulty case (no Byzantine agents).

2) **Case of $T > 1$:** We now generalize Theorem 1 to the case $T > 1$, i.e., each agent implements more than one local
Fig. 1. Plots show the error $\|\bar{x}_k - x^*\|^2$ in iteration $k = 1, \ldots, 120$ of local GD (cf. Algorithm 1) with four different aggregation schemes: averaging, CE, multi-KRUM, CWTM, and coordinatewise median. The benchmark corresponds to the fault-free execution of local GD. Solid lines show the mean performances of the schemes, and the shadows show the variance of their performances, observed over 100 runs. In the clockwise order, $f = 8$, $12$, $16$, and $20$. We observe that the accuracy improves for all the schemes with fewer Byzantine agents, and the CE filter performs consistently better than other schemes.

GD steps. This is a common practice in federated optimization. When $T > 1$, by (5) we have $\forall i \in \mathcal{H}$ and $t \in [0, T)$

$$x^i_{k,t+1} = \bar{x}_k - \alpha_k \sum_{\ell=0}^t \nabla q^i(x^i_{k,\ell}).$$

(14)

**Theorem 2:** Assume that (11) holds true and let $\alpha_k$ satisfy

$$\alpha_k = \alpha \leq \frac{\mu}{16TL^2}.$$  

(15)

Then we have

$$\|\bar{x}_k - x^*_{\mathcal{H}}\|^2 \leq \left(1 - \frac{\mu T \alpha}{6}\right)^k \|\bar{x}_0 - x^*_{\mathcal{H}}\|^2.$$  

(16)

**B. Stochastic Settings**

We next consider the setting where each agent only has access to the samples of its gradient, i.e., $G^i(\cdot) = \nabla Q^i(\cdot, X^i)$, where $X^i$ is a sequence of random variables sampled i.i.d from $\pi^i$. In the sequel, we denote by

$$\mathcal{P}_{k,t} = \cup_{i \in \mathcal{H}} \{\bar{x}_0, \ldots, \bar{x}_k, x^i_{k,1}, \ldots, x^i_{k,t}\}$$

the filtration containing all the history generated by Algorithm 1 up to time $k + t$. To study the convergence of Algorithm 1 we consider the following assumptions, which is often assumed in the literature of stochastic federated optimization [6].

**Assumption 3:** The random variables $X^i_k$, for all $i$ and $k$, are i.i.d., and there exists a positive constant $\sigma$ such that

$$\mathbb{E}[\nabla Q^i(x, X^i_k) | \mathcal{P}_{k,t}] = \nabla q^i(x) \ \forall x \in \mathbb{R}^d,$$

$$\mathbb{E}[\|\nabla Q^i(x, X^i_k) - \nabla q^i(x)\|^2 | \mathcal{P}_{k,t}] \leq \sigma^2 \ \forall x \in \mathbb{R}^d.$$  

Recall that $|B_k| + |\mathcal{H}_k| = |\mathcal{F}_k| = |\mathcal{H}|$, for any $k \geq 0$. Finally, for convenience we denote by

$$\nabla Q^i(x; X) = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \nabla Q^i(x; X^i)$$

where $X = (X^1, \ldots, X^{|\mathcal{H}|})^T$. By (5), for all $i \in \mathcal{H}$ and $t \in [0, T)$, we have

$$x^i_{k,t+1} = \bar{x}_k - \alpha_k \sum_{\ell=0}^t \nabla Q^i(x^i_{k,\ell}; X^i_{k,\ell}).$$  

(17)

**Theorem 3:** Suppose that Assumption 3 and condition (11) hold true. Moreover, let $\alpha_k$ be chosen as

$$\alpha_k = \alpha \leq \frac{\mu}{144TL^2}.$$  

(18)
Then we have
\[
E[\|\bar{x}_k - x^*_k\|^2] \leq \left(1 - \frac{\mu T \alpha}{18}\right)^k E[\|\bar{x}_0 - x^*_k\|^2] + \frac{162\sigma^2 T}{\mu} + \frac{54\sigma^2 f}{\mu L |\mathcal{H}|}. \tag{19}
\]

**Remark 3:** In Theorem 3 due to the constant step size, the mean square error converges linearly only to a ball centered at the origin. The size of this ball depends on two terms: 1) one depends on the step size \(\alpha\) which is often seen in the convergence of local GD with nonfaulty agents and 2) the other depends on the level of the gradient noise (or \(\sigma\)). The latter is due to the impact of the Byzantine agents and the stochastic gradient samples. Indeed, our comparative filter is designed to remove the potential bad values sent by the Byzantine agents, but not the variance of their stochastic samples. A potential solution for this issue is to let each agent sample a mini-batch of size \(m\), implying that \(\sigma^2\) in (19) is replaced by \(\sigma^2/m\). Thus, one can choose \(m\) large enough so that the mean square error can get arbitrarily close to zero. Finally, when \(\alpha_k \sim 1/k\), the convergence rate is sublinear \(O(1/k)\).

### IV. Experiments

To evaluate the efficacy of our proposed scheme, we simulate the problem of robust mean estimation in the federated framework. This problem serves as a test-case to empirically compare our scheme with others of similar computational costs, namely, multi-KRUM [17], CWTM [15], [22], and coordinate-wise median [22], [49]. For our experiments, we consider \(N = 50\) agents and varying number of Byzantine faulty agents. Each nonfaulty agent \(i\) has 100 noisy observations of a 10-D vector \(x^i\) with all elements of unit value. In particular, the sample set \(\mathcal{X}\) comprises 100 uniformly distributed samples with each sample \(X^i = x^i + Z^i\), where \(Z^i \sim \mathcal{N}(0, I_d)\), and \(Q^i(x; X^i) = (1/2)||x - X^i||^2\). In this case, \(x^*\) is the unique solution to problem (3) for any set of honest agents \(\mathcal{H}\). In our experimental settings, a Byzantine faulty agent \(j\) behaves like an honest agent with 100 uniformly distributed samples, however, each of its sample \(X^j = 2x^* + Z^j\), where \(Z^j \sim \mathcal{N}(0, I_d)\). That is, honest agents and Byzantine agents send information corresponding to the Gaussian noisy observations of \(x^*\) and \(2x^*\), respectively.

We simulate the stochastic setting of local GD (cf. Algorithm 1) with different number of faulty agents \(f \in \{8, 12, 16, 20, 24\}\), different values of \(T \in \{1, 2\}\), and different aggregation schemes in Step 2: CE, multi-KRUM, CWTM, coordinate-wise median, and simple averaging. The step-size \(\alpha_k = 0.1\) \(\forall k\). Each setting is run 100 times, and the observed errors \(\|\bar{x}_k - x^*\|^2\) are shown in Figs. 1 and 2.

### V. Conclusion

As suggested from our theoretical results, the final error upon using CE aggregation scheme decreases with the fraction of Byzantine faulty agents. We observe that CE aggregation scheme performs consistently better than multi-KRUM, CWTM, and median. Moreover, we also observe that increasing the number of local GD steps, i.e., \(T\), improves the fault-tolerance of CE aggregation scheme. However, the same cannot be said for other schemes.

### APPENDIX

#### A. Proof of Theorem 1

**Proof:** Using (7) and (10) we have
\[
\bar{x}_{k+1} = \frac{1}{|\mathcal{F}_k|} \sum_{i \in \mathcal{F}_k} x_{k,1}^i
\]
\[
= \frac{1}{|\mathcal{H}|} \left[ \sum_{i \in \mathcal{H}} x_{k,1}^i + \sum_{i \in \mathcal{B}_k} x_{k,1}^i - \sum_{i \in \mathcal{H}\setminus\mathcal{H}_k} x_{k,1}^i \right]
\]
\[
= \bar{x}_k - \frac{\alpha_k}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \nabla q^i(\bar{x}_k) + \frac{1}{|\mathcal{H}|} \left[ \sum_{i \in \mathcal{B}_k} x_{k,1}^i - \sum_{i \in \mathcal{H}\setminus\mathcal{H}_k} x_{k,1}^i \right]
\]
\[
= \bar{x}_k - \frac{\alpha_k}{|\mathcal{H}|} \nabla q^\mathcal{H}(\bar{x}_k) + \frac{1}{|\mathcal{H}|} \left[ \sum_{i \in \mathcal{B}_k} (x_{k,1}^i - \bar{x}_k) - \sum_{i \in \mathcal{H}\setminus\mathcal{H}_k} (x_{k,1}^i - \bar{x}_k) \right]
\]
\[
\tag{20}
\]
where the last equality is due to \(|\mathcal{B}_k| = |\mathcal{H}\setminus\mathcal{H}_k|\). Using the preceding relation we consider
\[ \| \bar{x}_{k+1} - x^*_H \|^2 = \| \bar{x}_k - x^*_H - \alpha_k \nabla q^H(\bar{x}_k) \|^2 \]

\[ + \left\| \frac{1}{| \mathcal{H} |} \sum_{i \in B_k} (x^*_k - \bar{x}_k) - \frac{1}{| \mathcal{H} |} \sum_{i \in \mathcal{H} \setminus B_k} (x^*_k - \bar{x}_k) \right\|^2 \]

\[ + \frac{2}{| \mathcal{H} |} \sum_{i \in B_k} (\bar{x}_k - x^*_H - \alpha_k \nabla q^H(\bar{x}_k))^T (x^*_{k,1} - \bar{x}_k) \]

\[ - \frac{2}{| \mathcal{H} |} \sum_{i \in \mathcal{H} \setminus B_k} (\bar{x}_k - x^*_H - \alpha_k \nabla q^H(\bar{x}_k))^T (x^*_{k,1} - \bar{x}_k). \] (21)

We next analyze each term on the right-hand side (RHS) of (21).

First, using (10) we have for all \( i \in \mathcal{H} \)

\[ x^*_{k,1} - \bar{x}_k = -\alpha_k \nabla q^H(\bar{x}_k). \]

In addition, under 2f-redundancy we have \( x^*_H \in \mathcal{A}^*_f \), where \( \mathcal{A}^*_f \) is the set of minimizers of \( q(x^*_H) \). This implies that \( \nabla q^H(x^*_H) = 0 \).

In addition, Assumption 1 implies that \( \nabla q^H \) is also L-Lipschitz continuous. Recall that \( | B_k | = | \mathcal{H} \setminus H_k | \). Then, by Assumption 1 we consider the last term on the RHS of (21)

\[ \frac{2}{| \mathcal{H} |} \sum_{i \in \mathcal{H} \setminus B_k} (\bar{x}_k - x^*_H - \alpha_k \nabla q^H(\bar{x}_k))^T (x^*_{k,1} - \bar{x}_k) \]

\[ = \frac{2\alpha_k}{| \mathcal{H} |} \sum_{i \in \mathcal{H} \setminus B_k} (\bar{x}_k - x^*_H - \alpha_k \nabla q^H(\bar{x}_k))^T \nabla q^H(\bar{x}_k) \]

\[ = \frac{2\alpha_k}{| \mathcal{H} |} \sum_{i \in \mathcal{H} \setminus B_k} (\bar{x}_k - x^*_H - \alpha_k (\nabla q^H(\bar{x}_k) - \nabla q^H(x^*_H)))^T \]

\[ \times (\nabla q^H(\bar{x}_k) - \nabla q^H(x^*_H)). \]

\[ \leq \frac{2\alpha_k}{| \mathcal{H} |} \sum_{i \in \mathcal{H} \setminus B_k} (\| \bar{x}_k - x^*_H \| + \alpha_k \| \nabla q^H(\bar{x}_k) \|) (\| x^*_{k,1} - \bar{x}_k \|) \]

\[ = \alpha_k \| \nabla q^H(\bar{x}_k) - \nabla q^H(x^*_H) \| \leq L \alpha_k \| \bar{x}_k - x^*_H \|. \] (23)

Second, using our CE filter [a.k.a. (6)] and (10) there exists \( j \in \mathcal{H} \) such that for all \( i \in B_k \) we have

\[ \| x^*_{k,1} - \bar{x}_k \| \leq \| x^*_{k,1} - \bar{x}_k \| = \alpha_k \| \nabla q^H(\bar{x}_k) \| \]

\[ = \alpha_k \| \nabla q^H(\bar{x}_k) - \nabla q^H(x^*_H) \| \leq L \alpha_k \| \bar{x}_k - x^*_H \|. \]

Next, using (10) and (23) we obtain

\[ \left\| \frac{1}{| \mathcal{H} |} \sum_{i \in B_k} (x^*_{k,1} - \bar{x}_k) - \frac{1}{| \mathcal{H} |} \sum_{i \in \mathcal{H} \setminus B_k} (x^*_{k,1} - \bar{x}_k) \right\|^2 \]

\[ \leq \frac{2 | B_k |}{| \mathcal{H} |^2} \left( \sum_{i \in B_k} \| x^*_{k,1} - \bar{x}_k \|^2 + \sum_{i \in \mathcal{H} \setminus B_k} \| x^*_{k,1} - \bar{x}_k \|^2 \right) \]

\[ \leq 4L^2 | B_k |^2 \alpha^2_k \| \bar{x}_k - x^*_H \|^2. \] (25)

Finally, using Assumption 2 we obtain

\[ \| \bar{x}_k - x^*_H - \alpha_k \nabla q^H(\bar{x}_k) \|^2 = \| \bar{x}_k - x^*_H \|^2 \]

\[ - 2 \alpha_k \nabla q^H(\bar{x}_k)^T (\bar{x}_k - x^*_H) + \alpha_k^2 \| \nabla q^H(\bar{x}_k) - \nabla q^H(x^*_H) \|^2 \]

\[ \leq (1 - 2 \alpha_k + L^2 \alpha_k^2) \| \bar{x}_k - x^*_H \|^2. \] (26)

Substituting (22)–(26) into (21), we obtain

\[ \| \bar{x}_{k+1} - x^*_H \|^2 \leq (1 - 2 \alpha_k) \| \bar{x}_k - x^*_H \|^2 \]

\[ + \frac{4L^2 | B_k |^2}{| \mathcal{H} |^2} \alpha^2_k \| \bar{x}_k - x^*_H \|^2 \]

\[ + \left( \frac{2L | B_k |}{| \mathcal{H} |} + \frac{2L^2 | B_k |^2}{| \mathcal{H} |^2} \right) \| \bar{x}_k - x^*_H \|^2 \]

\[ + \left( \frac{2L | B_k |}{| \mathcal{H} |} + \frac{2L^2 | B_k |^2}{| \mathcal{H} |^2} \right) \| \bar{x}_k - x^*_H \|^2 \]

\[ \leq (1 - 2 \mu \alpha_k + L^2 \alpha_k^2) \| \bar{x}_k - x^*_H \|^2 \]

\[ + \left( 1 + \frac{4f}{| \mathcal{H} |} \right) \frac{4f^2}{| \mathcal{H} |^2} \frac{L^2 \alpha_k^2}{| \mathcal{H} |} \| \bar{x}_k - x^*_H \|^2 \]

\[ \leq \left( 1 - 2 \mu \alpha_k + L^2 \alpha_k^2 \right) \| \bar{x}_k - x^*_H \|^2 \]

\[ + \left( 1 + \frac{4f}{| \mathcal{H} |} \right) \left( \frac{4f^2}{| \mathcal{H} |^2} L^2 \alpha_k^2 \right) \| \bar{x}_k - x^*_H \|^2 \] (27)

where the last inequality we use \(| B_k | \leq f \). Using (11) gives \( \mu - \frac{2Lf}{| \mathcal{H} |} \geq \mu - \frac{2\mu}{6} = \frac{1}{6} \), which when substituting into (27) and using (12) gives (13)

\[ \| \bar{x}_{k+1} - x^*_H \|^2 \leq \left( 1 - \frac{2\alpha_k}{3} \right) \| \bar{x}_k - x^*_H \|^2 \]

\[ + \left( 1 + \frac{4f}{| \mathcal{H} |} + \frac{4f^2}{| \mathcal{H} |^2} \right) \frac{L^2 \alpha_k^2}{| \mathcal{H} |} \| \bar{x}_k - x^*_H \|^2 \]

\[ \leq \left( 1 - \frac{2\mu}{3} \right) \| \bar{x}_k - x^*_H \|^2 + 3L^2 \alpha_k^2 \| \bar{x}_k - x^*_H \|^2 \]

\[ \leq \left( 1 - \frac{\mu \alpha_k}{6} \right) \| \bar{x}_k - x^*_H \|^2 \leq \left( \frac{1}{1/3} \right) \| \bar{x}_0 - x^*_H \|^2 \]

where the second inequality is due to \( f / | \mathcal{H} | \leq 1/3 \) and the third inequality is due to \( \alpha_k = \alpha \leq \mu / (4L^2) \).
B. Proof of Theorem 2

Proof: Due to $T > 1$, the proof of Theorem 2 is different to the one in Theorem 1 at the way we quantify the size of $|x_{k,t} - \bar{x}_k|_t$ for any $t \in [0, T]$. Thus, to show (16) we first provide an upper bound of this quantity. Note that we do not assume the gradient being bounded. By (15) we have

$$LT \alpha_k = LT \alpha \leq \frac{\mu}{16L} \leq \frac{1}{16} \leq \ln(2).$$

By (5), Assumption 1, and $\nabla q^i(x_{k,t}) = 0$ we have for all $t \in [0, T - 1]$ and $i \in \mathcal{H}$

$$\|x_{k,t+1}^i - x_{k,t}^i\| = \alpha_k \|\nabla q^i(x_{k,t})\| \leq L \alpha_k \|x_{k,t}^i - x_{k,t-1}^i\|$$

which using $1 + x \leq \exp(x)$ for $x > 0$ implies

$$\|x_{k,t+1}^i - x_{k,t}^i\| \leq (1 + L \alpha_k) \|x_{k,t}^i - x_{k,t-1}^i\|$$

$$\leq \exp(L \alpha_k) \|x_{k,t}^i - x_{k,t-1}^i\| \leq \exp(LT \alpha_k) \|x_{k,0}^i - x_{k,T}^i\|$$

$$\leq 2 \|x_{k,T}^i - x_{k,T}^i\|$$

where the last inequality we use $LT \alpha_k \leq \ln(2)$ and $x_{k,0}^i = \bar{x}_k$. Using the preceding two relations gives for all $t \in [0, T]$

$$\|x_{k,t+1}^i - x_{k,t}^i\| \leq L \alpha_k \|x_{k,t}^i - x_{k,t-1}^i\| \leq 2L \alpha_k \|x_{k,T}^i - x_{k,T}^i\|$$

Using this relation and $x_{k,0}^i = \bar{x}_k$ gives $\forall t \in [0, T - 1]$

$$\|x_{k,t+1}^i - \bar{x}_k\| = \frac{\sum_{\ell=0}^{T} x_{k,t+1}^i - x_{k,t}^i}{t} \leq \frac{2LT \alpha_k \|x_{k,T}^i - x_{k,T}^i\|}{t}$$

(28)

Next, we consider

$$\bar{x}_{k+1} = \frac{1}{|\mathcal{F}_k|} \sum_{i \in \mathcal{F}_k} x_{k,T}^i$$

$$= \frac{1}{|\mathcal{H}|} \left[ \sum_{i \in \mathcal{H}} x_{k,T}^i + \sum_{i \in \mathcal{B}_k} x_{k,T}^i - \sum_{i \in \mathcal{H} \setminus \mathcal{H}_k} x_{k,T}^i \right]$$

(14)

$$= \bar{x}_k - \frac{\alpha_k}{|\mathcal{H}|} \sum_{i \in \mathcal{H} \setminus \mathcal{H}_k} \nabla q^i(x_{k,t}^i)$$

$$+ \frac{1}{|\mathcal{H}|} \left[ \sum_{i \in \mathcal{B}_k} x_{k,T}^i - \sum_{i \in \mathcal{H} \setminus \mathcal{H}_k} x_{k,T}^i \right]$$

$$= \bar{x}_k - T \alpha_k \nabla q^H(\bar{x}_k)$$

$$- \frac{\alpha_k}{|\mathcal{H}|} \sum_{i \in \mathcal{H} \setminus \mathcal{H}_k} \left( \nabla q^i(x_{k,t}^i) - \nabla q^i(\bar{x}_k) \right)$$

$$+ \frac{1}{|\mathcal{H}|} \left[ \sum_{i \in \mathcal{B}_k} x_{k,T}^i - \sum_{i \in \mathcal{H} \setminus \mathcal{H}_k} x_{k,T}^i \right]$$

(29)

where the last equality is due to $|\mathcal{B}_k| = |\mathcal{H} \setminus \mathcal{H}_k|$. For convenience, we denote by

$$V_k^e = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{B}_k} (x_{k,T}^i - \bar{x}_k) - \sum_{i \in \mathcal{H} \setminus \mathcal{H}_k} (x_{k,T}^i - \bar{x}_k)$$

Using (29) gives

$$\|\bar{x}_{k+1} - x_{k,T}^i\|^2 = \|\bar{x}_k - x_{k,T}^i - T \alpha_k \nabla q^H(\bar{x}_k)\|^2 + \|V_k^e\|^2$$

$$+ \frac{\alpha_k}{|\mathcal{H}|} \sum_{i \in \mathcal{H} \setminus \mathcal{H}_k} \sum_{\ell=0}^{T-1} \left( \nabla q^i(x_{k,t}^i) - \nabla q^i(\bar{x}_k) \right)$$

$$- \frac{2\alpha_k}{|\mathcal{H}|} \sum_{i \in \mathcal{H} \setminus \mathcal{H}_k} \sum_{\ell=0}^{T-1} (x_{k,T}^i - x_{k,T}^i)^T \left( \nabla q^i(x_{k,t}^i) - \nabla q^i(\bar{x}_k) \right)$$

$$+ \frac{2T \alpha_k^2}{|\mathcal{H}|} \sum_{i \in \mathcal{H} \setminus \mathcal{H}_k} \sum_{\ell=0}^{T-1} \nabla q^i(\bar{x}_k)^T \left( \nabla q^i(x_{k,t}^i) - \nabla q^i(\bar{x}_k) \right)$$

$$+ \frac{2}{T} \left( \bar{x}_k - x_{k,T}^i - T \alpha_k \nabla q^H(\bar{x}_k) \right)^T V_k^e.$$

(30)

We next consider each term on the RHSs of (30). First, using Assumptions 1 and 2, and $\nabla q^H(\bar{x}_k) = 0$, we have

$$\|\bar{x}_k - x_{k,T}^i - T \alpha_k \nabla q^H(\bar{x}_k)\|^2$$

$$= \|\bar{x}_k - x_{k,T}^i\|^2 - 2T \alpha_k \left( \bar{x}_k - x_{k,T}^i \right)^T \nabla q^H(\bar{x}_k)$$

$$+ T^2 \alpha_k^2 \|\nabla q^H(\bar{x}_k)\|^2$$

$$= \|\bar{x}_k - x_{k,T}^i\|^2 - 2T \alpha_k \left( \bar{x}_k - x_{k,T}^i \right)^T \nabla q^H(\bar{x}_k)$$

$$+ T^2 \alpha_k^2 \|\nabla q^H(\bar{x}_k)\|^2$$

$$\leq (1 - 2\mu T \alpha_k + T^2 \alpha_k^2) \|\bar{x}_k - x_{k,T}^i\|^2.$$

(31)

Second, by (6), there exists $j \in \mathcal{H} \setminus \mathcal{H}_k$ such that $\|x_{k,T}^j - \bar{x}_k\| \leq \|x_{k,T}^j - \bar{x}_k\|_t$ for all $i \in \mathcal{B}_k$. Then we have

$$\|V_k^e\|^2 = \left\| \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{B}_k} x_{k,T}^i - \bar{x}_k - \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H} \setminus \mathcal{H}_k} x_{k,T}^i - \bar{x}_k \right\|^2$$

$$\leq \frac{2|\mathcal{B}_k|}{|\mathcal{H}|^2} \|x_{k,T}^i - \bar{x}_k\|^2 + \frac{2|\mathcal{B}_k|}{|\mathcal{H}|^2} \sum_{i \in \mathcal{H} \setminus \mathcal{H}_k} \|x_{k,T}^i - \bar{x}_k\|^2$$

$$\leq \frac{2|\mathcal{B}_k|}{|\mathcal{H}|^2} \|x_{k,T}^i - \bar{x}_k\|^2 + \frac{2|\mathcal{B}_k|}{|\mathcal{H}|^2} \sum_{i \in \mathcal{H} \setminus \mathcal{H}_k} \|x_{k,T}^i - \bar{x}_k\|^2$$

$$\leq \frac{\|\bar{x}_k - x_{k,T}^i\|^2}{16 T^2 \alpha_k^2 |\mathcal{B}_k|^2} \|\bar{x}_k - x_{k,T}^i\|^2.$$  

(32)
Third, using (28) yields

\[ \left\| \frac{\alpha_k}{|H|} \sum_{i \in \mathcal{H}} \sum_{t=0}^{T-1} (\nabla q^i(x_{k,t}^i) - \nabla q^i(\bar{x}_k)) \right\|^2 \leq \frac{L^2 T \alpha_k^2}{|H|} \sum_{i \in \mathcal{H}} \sum_{t=0}^{T-1} \|x_{k,t}^i - \bar{x}_k\|^2 \leq 4 L^2 T^2 \alpha_k^2 \| \bar{x}_k - x_H^* \|^2. \]

(33)

Fourth, using Assumption 1, we consider

\[ \frac{2 \alpha_k}{|H|} (\bar{x}_k - x_H^*)^T \sum_{t=0}^{T-1} \sum_{i \in \mathcal{H}} (\nabla q^i(x_{k,t}^i) - \nabla q^i(\bar{x}_k)) \leq \frac{2 L \alpha_k}{|H|} \sum_{t=0}^{T-1} \sum_{i \in \mathcal{H}} \|\bar{x}_k - x_{k,t}^i\| \|x_{k,t}^i - \bar{x}_k\| \leq 4 T^2 L^2 \alpha_k^2 \| \bar{x}_k - x_H^* \|^2. \]

(34)

Fifth, using Assumption 1 and \( \nabla q^H(x_H^*) = 0 \), we have

\[ \frac{2 T \alpha_k^2}{|H|} \sum_{i \in \mathcal{H}} \sum_{t=0}^{T-1} \|\nabla q^H(\bar{x}_k)\| \|\nabla q^H(\bar{x}_k)\| \|\bar{x}_k - x_{k,t}^i\| \leq \frac{4 L^2 T^3 \alpha_k^3}{|H|} \|\nabla q^H(\bar{x}_k)\| \|\bar{x}_k - x_H^*\|^2 \leq 4 L^2 T^3 \alpha_k^3 \| \bar{x}_k - x_H^* \|^2. \]

(35)

Next, by (6) and similar to (32), we have

\[ \|V_k\| = \left\| \frac{1}{|H|} \sum_{i \in \mathcal{B}_k} x_{k,t}^i - \bar{x}_k \right\| \leq \frac{1}{|H|} \sum_{i \in \mathcal{B}_k} x_{k,t}^i - \bar{x}_k \leq \frac{1}{|H|} \sum_{i \in \mathcal{H} \setminus \mathcal{H}_k} \|x_{k,t}^i - \bar{x}_k\| \leq \frac{2 T L |\mathcal{B}_k|}{|H|} \leq 2 T L |\mathcal{B}_k| \leq \frac{2 T L |\mathcal{B}_k|}{|H|} \alpha_k \| \bar{x}_k - x_H^* \| \]

(28)

which by using \( \nabla q^H(x_H^*) = 0 \), (6), and (33), we have

\[ \begin{align*}
2 (\bar{x}_k - x_H^* - T \alpha_k \nabla q^H(\bar{x}_k))^T V_k \\
\leq 2 \left( \left\| \nabla q^H(\bar{x}_k) - \nabla q^H(\bar{x}_k) \right\| \|x_H^* - \bar{x}_k\| + \nabla q^H(\bar{x}_k) \right) \left\| V_k \right\| \leq \frac{4 T L |\mathcal{B}_k|}{|H|} \left( 1 + LT \alpha_k \right) \alpha_k \| \bar{x}_k - x_H^* \|^2.
\end{align*} \]

(36)

Substituting (31)–(36) into (30) and using \(|\mathcal{B}_k| \leq f\) yields

\[ \|x_{k+1} - x_H^*\|^2 \leq \left( 1 - 2 \mu T \alpha_k + T^2 L^2 \alpha_k^2 \right) \| \bar{x}_k - x_H^* \|^2 + \frac{16 T^2 L^2 f^2}{|H|^2} \alpha_k^2 \| \bar{x}_k - x_H^* \|^2 + 4 L^2 T^4 \alpha_k^4 \| \bar{x}_k - x_H^* \|^2. \]

(37)

By (11), we have \( \mu \leq \frac{2 L T}{|H|} \geq \frac{\mu}{3} \leq \frac{\mu}{16 T L} \) and \( \alpha_k \) is bounded hence bounded. Using the preceding two relations into (16) and recalling that \( \alpha_k = \alpha \leq \mu/(16 T L) \), we obtain

\[ \|x_{k+1} - x_H^*\|^2 \leq \left( 1 - \frac{\mu T \alpha}{3} \right) \| \bar{x}_k - x_H^* \|^2 + 8 L^2 T^2 \alpha^2 \| \bar{x}_k - x_H^* \|^2 \]

\[ \leq \left( 1 - \frac{\mu T \alpha}{6} \right) \| \bar{x}_k - x_H^* \|^2 \leq \left( 1 - \frac{\mu T \alpha}{6} \right)^{k+1} \| \bar{x}_0 - x_H^* \|^2 \]

which concludes our proof.

\[ \blacksquare \]

**C. Proof of Theorem 3**

**Proof:** Using (5), we have for all \( t \in [0, T - 1] \) and \( i \in \mathcal{H} \)

\[ \|x_{k,t+1}^i - x_H^* - x_{k,t}^i\| \leq \alpha_k \| \nabla q^i(x_{k,t}^i) \| \| \bar{x}_k - x^*_H \| \leq \alpha_k \| \nabla q^i(x_{k,t}^i) \| + \alpha_k \| \nabla q^i(x_{k,t}^i) - \nabla q^i(x_{k,t}^i) \| \]

which by using Assumptions 1 and 3, and \( \nabla q^i(x_H^*) = 0 \) yields

\[ \begin{align*}
\mathbb{E}[\|x_{k,t+1}^i - x_H^* - x_{k,t}^i\| + \|x_{k,t}^i - x_H^*\| \|P_{k,t}\|] \\
\leq \mathbb{E}[\|x_{k,t+1}^i - x_{k,t}^i\|] + \mathbb{E}[\|P_{k,t}\|] \\
\leq \alpha_k \mathbb{E}[\| \nabla q^i(x_{k,t}^i) - \nabla q^i(x_{k,t}^i) \| + \alpha_k \mathbb{E}[\| \nabla q^i(x_{k,t}^i) - \nabla q^i(x_{k,t}^i) \| \|P_{k,t}\|)] \\
\leq \alpha_k \mathbb{E}[\| \nabla q^i(x_{k,t}^i) - x_H^* \| + \alpha_k].
\end{align*} \]

Using \( 1 + x \leq \exp(x) \) for \( x > 0 \), the preceding relation gives

\[ \mathbb{E}[\|x_{k,t+1}^i - x_H^*\|] \leq (1 + \alpha_k) \mathbb{E}[\|x_{k,t}^i - x_H^*\| + \alpha_k] \\
\leq (1 + \alpha_k) \mathbb{E}[\|x_{k,0}^i - x_H^*\| + \alpha_k \sum_{\ell=0}^{t-1} (1 + \alpha_k)^{t-1-\ell}] + \alpha_k \sum_{\ell=0}^{t-1} (1 + \alpha_k)^{-1-\ell} \]

\[ = \exp(Lt \alpha_k) \mathbb{E}[\|x_{k,0}^i - x_H^*\|] + \alpha_k \sum_{\ell=0}^{t-1} (1 + \alpha_k)^{-1-\ell}. \]
\[ \leq \exp(L\alpha_k)E[\|x_{k,0}^i - x_{H}^i\|] + \frac{\sigma \alpha_k \exp(L\alpha_k)}{L}\]

\[ \leq 2E[\|\bar{x}_k - x_{\star H}^i\|] + \frac{2\sigma}{T} \]

where the last inequality we use \(L\alpha_k \leq \ln(2)\) and \(x_{k,0}^i = \bar{x}_k\). Thus, we obtain for all \(t \in [0, T - 1]\)

\[ E[\|x_{k,t+1}^i - x_{H}^i\|] \leq L\alpha_k E[\|x_{k,t}^i - x_{H}^i\|] + \sigma \alpha_k \]

\[ \leq 2L\alpha_k E[\|\bar{x}_k - x_{\star H}^i\|] + 3\sigma \alpha_k \]

which, by using \(x_{k,0}^i = \bar{x}_k\), gives \(\forall t \in [0, T - 1]\) and \(i \in H\)

\[ \begin{align*}
E[\|x_{k,t+1}^i - \bar{x}_k\|] &= E\left[ \left\| \sum_{\ell=0}^{t} x_{k,\ell+1}^i - x_{k,t}^i \right\| \right] \\
&\leq \sum_{\ell=0}^{t} E[\|x_{k,\ell+1}^i - x_{k,\ell}^i\|] \\
&\leq 2L\alpha_k E[\|\bar{x}_k - x_{\star H}^i\|] + 3\sigma \alpha_k \quad (38)
\end{align*} \]

Similarly, we consider \(i \in H\)

\[ E[\|x_{k,t+1}^i - x_{H}^i\|] = E[\|x_{k,t}^i - x_{\star H}^i - \alpha_k \nabla q^i(x_{k,t}^i; X_{k,t}^i)\|] \]

\[ = E[\|x_{k,t}^i - x_{\star H}^i\|^2] - 2E[\langle x_{k,t}^i - x_{\star H}^i, \nabla q^i(x_{k,t}^i; X_{k,t}^i) \rangle] + E[\|\alpha_k \nabla q^i(x_{k,t}^i; X_{k,t}^i)\|^2] \]

\[ \leq (1 + L^2 \alpha_k^2) E[\|x_{k,t}^i - x_{\star H}^i\|^2] + \sigma \alpha_k^2 \]

where the last inequality we use the convexity of \(q^i\), Assumptions 1 and 3. Using the relation \(1 + x \leq \exp(x)\) for all \(x \geq 0\) and \(x_{k,0}^i = \bar{x}_k\), the preceding relation gives for all \(t \in [0, T - 1]\)

\[ E[\|x_{k,t+1}^i - x_{H}^i\|] \leq (1 + L^2 \alpha_k^2) E[\|\bar{x}_k - x_{\star H}^i\|^2] \]

\[ + \sigma \alpha_k^2 \sum_{\ell=0}^{t-1} (1 + L^2 \alpha_k^2)^{t-\ell} \leq \exp(L^2 T \alpha_k^2) E[\|\bar{x}_k - x_{\star H}^i\|^2] \]

\[ + \frac{\sigma^2 (1 + L^2 \alpha_k^2)}{L^2} \leq 2E[\|\bar{x}_k - x_{\star H}^i\|^2] + \frac{2\sigma^2}{L^2} \]

where the last inequality is due to \(\exp(L^2 T \alpha_k^2) \leq 2\). Using the relation above we obtain for all \(t \in [0, T - 1]\) and \(i \in H\)

\[ E[\|x_{k,t+1}^i - \bar{x}_k\|^2] = E\left[ \left\| \sum_{\ell=0}^{t} (x_{k,\ell+1}^i - x_{k,\ell}^i) \right\|^2 \right] \]

\[ \leq \sum_{\ell=0}^{t} E[\|x_{k,\ell+1}^i - x_{k,\ell}^i\|^2] = t \alpha_k^2 \sum_{\ell=0}^{t} E[\|\nabla q^i(x_{k,\ell}^i; X_{k,\ell}^i)\|^2] \]

\[ = t \alpha_k^2 \sum_{\ell=0}^{t} E[\|\nabla q^i(x_{k,\ell}^i; X_{k,\ell}^i) - q^i(x_{k,\ell}^i)\|^2] \]

\[ + t \alpha_k^2 \sum_{\ell=0}^{t} \frac{\sigma}{L^2} \sum_{\ell=0}^{t} E[\|\nabla q^i(x_{k,\ell}^i; X_{k,\ell}^i)\|^2] \]

\[ \leq T \alpha_k^2 \sum_{\ell=0}^{t} E[\|x_{k,\ell}^i - x_{\star H}^i\|^2] \]

\[ \leq T^2 \alpha_k^2 \sum_{\ell=0}^{t} E[\|x_{k,\ell}^i - x_{\star H}^i\|^2] \]

\[ \leq 2L^2 T^2 \alpha_k^2 E[\|\bar{x}_k - x_{\star H}^i\|^2] + 3T^2 \sigma^2 \alpha_k^2 \quad (39) \]

Note that we have \(|F_k| = |H| = N - f\). We then consider

\[ \bar{x}_{k+1} = \frac{1}{|F_k|} \sum_{i \in F_k} x_{k,t}^i \]

\[ = \frac{1}{|H|} \left[ \sum_{i \in H} x_{k,t}^i + \sum_{i \in B} x_{k,t}^i - \sum_{i \in H \setminus H_k} x_{k,t}^i \right] \]

\[ = \bar{x}_k - \frac{\alpha_k}{|H|} \sum_{i \in H} \sum_{t=0}^{T-1} \nabla q^i(x_{k,t}^i; X_{k,t}^i) \]

\[ - \frac{\alpha_k}{|H|} \sum_{i \in B} \sum_{t=0}^{T-1} \nabla q^i(x_{k,t}^i; X_{k,t}^i) \]

\[ + \frac{1}{|H|} \left[ \sum_{i \in B} (x_{k,t}^i - \bar{x}_k) - \sum_{i \in H \setminus H_k} (x_{k,t}^i - \bar{x}_k) \right] \]

For convenience, we denote by

\[ V_k^e = \frac{1}{|H|} \sum_{i \in B} (x_{k,t}^i - \bar{x}_k) - \sum_{i \in H \setminus H_k} (x_{k,t}^i - \bar{x}_k). \]

Using (40), we consider

\[ \|\bar{x}_{k+1} - x_{\star H}^i\|^2 = \|\bar{x}_k - x_{\star H}^i - \alpha_k \nabla q^i(\bar{x}_k)\|^2 + V_k^e \]

\[ \leq \frac{\alpha_k^2}{|H|} \sum_{i \in H} \sum_{t=0}^{T-1} \nabla q^i(x_{k,t}^i; X_{k,t}^i) - \nabla q^i(\bar{x}_k) \]

\[ + \frac{2\alpha_k}{|H|} \sum_{i \in B} \sum_{t=0}^{T-1} \nabla q^i(x_{k,t}^i; X_{k,t}^i) - \nabla q^i(\bar{x}_k) \]

\[ + 2 \left( \bar{x}_k - x_{\star H}^i - \alpha_k \nabla q^H(\bar{x}_k) \right)^T V_k^e \]

\[ + 2 \left( \bar{x}_k - x_{\star H}^i - \alpha_k \nabla q^H(\bar{x}_k) \right)^T V_k^e \]

\[ + 2 \left( \bar{x}_k - x_{\star H}^i - \alpha_k \nabla q^H(\bar{x}_k) \right)^T V_k^e \]

\[ + 2 \left( \bar{x}_k - x_{\star H}^i - \alpha_k \nabla q^H(\bar{x}_k) \right)^T V_k^e \]

\[ \leq \left( 1 - 2\mu \alpha_k + T^2 \alpha_k^2 \right) \|\bar{x}_k - x_{\star H}^i\|^2. \]
Second, by (6) there exists \( j \in \mathcal{H} \setminus \mathcal{H}_k \) such that \( \| x_{k,T}^j - \bar{x}_k \| \leq \| x_{k,T}^j - \bar{x}_k \| \) for all \( i \in \mathcal{B}_k \). Then we have

\[
\mathbb{E} \left[ \left\| V_k^x \right\|^2 \right] = \frac{2 |B_k|}{|\mathcal{H}|^2} \left( \sum_{i \in \mathcal{B}_k} \mathbb{E} \left[ \| x_{k,T}^i - \bar{x}_k \|^2 \right] + \sum_{i \in \mathcal{H} \setminus \mathcal{H}_k} \mathbb{E} \left[ \| x_{k,T}^i - \bar{x}_k \|^2 \right] \right) \leq 2 \frac{|B_k|}{|\mathcal{H}|^2} \left( \sum_{i \in \mathcal{B}_k} \mathbb{E} \left[ \| x_{k,T}^i - \bar{x}_k \|^2 \right] + \sum_{i \in \mathcal{H} \setminus \mathcal{H}_k} \mathbb{E} \left[ \| x_{k,T}^i - \bar{x}_k \|^2 \right] \right) \leq 2 \frac{8T^2 L^2 |B_k|^2}{|\mathcal{H}|^2} \mathbb{E} \left[ \| x_k - x_h^* \|^2 \right] + \frac{12T^2 \sigma^2 |B_k|^2}{|\mathcal{H}|^2} \alpha_k^2. \] (43)

Third, using Assumptions 1 and 3, and (39) yields

\[
\mathbb{E} \left[ \left\| \frac{\alpha_k}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \sum_{t=0}^{T-1} (\nabla Q_i^H(x_{k,t}^i; X_{k,t}^i) - \nabla q_i^H(\bar{x}_k)) \right\|^2 \right] \leq \frac{T^2 \sigma^2 \alpha_k^2}{|\mathcal{H}|} + \frac{2T^2 \alpha_k^2}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \sum_{t=0}^{T-1} \mathbb{E} \left[ \| x_{k,t}^i - \bar{x}_k \|^2 \right] \leq 2T^4 \mathbb{E} \left[ \| x_k - x_h^* \|^2 \right] + 3L^2 \sigma^2 T^2 \alpha_k^2. \] (44)

Similarly, we consider

\[
-2 \mathbb{E} \left[ (\bar{x}_k - x_h^*)^T (\nabla Q_i^H(x_{k,t}^i; X_{k,t}^i) - \nabla q_i^H(\bar{x}_k)) \right] | \mathcal{H}_k \leq 2 |B_k| \alpha_k \mathbb{E} \left[ \left\| \nabla Q_i^H(x_{k,t}^i; X_{k,t}^i) - \nabla q_i^H(\bar{x}_k) \right\|^2 \right] \leq 2L^2 \mathbb{E} \left[ \left\| x_k - x_h^* \right\|^2 + \frac{1}{\alpha_k} \right] \left\| x_k^i - \bar{x}_k \right\|^2 \] (39)

we obtain

\[
-2 \mathbb{E} \left[ (\bar{x}_k - x_h^*)^T (\nabla Q_i^H(x_{k,t}^i; X_{k,t}^i) - \nabla q_i^H(\bar{x}_k)) \right] \leq L^2 \alpha_k \mathbb{E} \left[ \left\| \bar{x}_k - x_h^* \right\|^2 + \left\| x_k - x_h^* \right\|^2 \right] + 3 \sigma^2 \alpha_k \mathbb{E} \left[ \left\| x_k - x_h^* \right\|^2 \right] \leq 3L^2 \alpha_k \mathbb{E} \left[ \left\| \bar{x}_k - x_h^* \right\|^2 + 3 \sigma^2 \alpha_k \mathbb{E} \left[ \left\| x_k - x_h^* \right\|^2 \right]. \] (45)

Using the preceding relation we obtain an upper bound for the fourth term on the RHS of (41)

\[
-2 \mathbb{E} \left[ \frac{2\alpha_k}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \sum_{t=0}^{T-1} (\bar{x}_k - x_h^*)^T (\nabla Q_i^H(x_{k,t}^i; X_{k,t}^i) - \nabla q_i^H(\bar{x}_k)) \right] \leq 3L^2 \alpha_k \mathbb{E} \left[ \left\| \bar{x}_k - x_h^* \right\|^2 \right] + 3 \sigma^2 \alpha_k \mathbb{E} \left[ \left\| x_k - x_h^* \right\|^2 \right]. \] (46)

Similarly, we provide an upper bound for the fifth term on the RHS of (41). Using Assumption 1 and \( \nabla q_i^H(x_h^*) = 0 \) gives

\[
2 \mathbb{E} \left[ \nabla q_i^H(\bar{x}_k)^T (\nabla Q_i^H(x_{k,t}^i; X_{k,t}^i) - \nabla q_i^H(\bar{x}_k)) \right] \leq 2 \mathbb{E} \left[ \nabla q_i^H(\bar{x}_k)^T (\nabla Q_i^H(x_{k,t}^i; X_{k,t}^i) - \nabla q_i^H(x_h^*)) \right] \leq 2L^2 \mathbb{E} \left[ \left\| x_k - x_h^* \right\|^2 \right] + 2L^2 \mathbb{E} \left[ \left\| x_k - x_h^* \right\|^2 \right] + 3L^2 \alpha_k \mathbb{E} \left[ \left\| \bar{x}_k - x_h^* \right\|^2 \right] + 3 \sigma^2 \alpha_k \mathbb{E} \left[ \left\| x_k - x_h^* \right\|^2 \right]. \] (46)

Finally, we analyze the last term on the RHS of (41). Using \( \nabla q_i^H(x_h^*) = 0 \), (6), (43), and the relation \( 2(x, y) \leq \eta \| x \|^2 + \| y \|^2 / \eta \) for any \( \eta > 0 \) and \( x, y \in \mathbb{R} \), taking the expectation on both sides of the preceding relation and using
where the last inequality we use $|B_k| \leq f$ and using $LT \alpha_k \leq \ln(2) \leq 1$ we obtain that

$$
E \left[ \|x_{k+1} - x^*_k\|^2 \right] \leq \left( 1 - 2 \left( \mu - \frac{17Lf}{6|H|} \right) T \alpha_k \right) E \left[ \|x_k - x^*_k\|^2 \right] + \frac{8T^2L^2f^2}{|H|^2} \alpha_k E \left[ \|x_k - x^*_k\|^2 \right] + \frac{12T^2\sigma^2f^2}{|H|^2} \alpha_k
$$

$$
+ 6T^2L^2\alpha_k E \left[ \|x_k - x^*_k\|^2 \right] + 7\sigma^2T^2\alpha_k + \frac{4T^2f}{L|H|} \alpha_k
$$

$$
+ \frac{3T^2L^2f\sigma^2}{|H|} E \left[ \|x_k - x^*_k\|^2 \right].
$$

Using (11) gives $\mu - \frac{17Lf}{6|H|} \geq \frac{\mu}{18}$. Thus, since $f/|H| \leq \frac{3}{8}$, we obtain from (48)

$$
E \left[ \|x_{k+1} - x^*_k\|^2 \right] \leq \left( 1 - \frac{\mu T \alpha_k}{9} \right) E \left[ \|x_k - x^*_k\|^2 \right] + \left( \frac{6 + 3f}{|H|^2} + \frac{8f^2}{|H|^2} \right) T^2L^2\alpha_k E \left[ \|x_k - x^*_k\|^2 \right]
$$

$$
+ 12T^2\sigma^2f^2 \alpha_k + 7\sigma^2T^2\alpha_k + \frac{3T^2\sigma^2f}{L|H|} \alpha_k
$$

$$
\leq \left( 1 - \frac{\mu T \alpha_k}{9} \right) E \left[ \|x_k - x^*_k\|^2 \right] + 8T^2L^2\alpha_k E \left[ \|x_k - x^*_k\|^2 \right] + 9\sigma^2T^2\alpha_k + \frac{3T^2\sigma^2f}{L|H|} \alpha_k.
$$

Since $\alpha_k = \alpha \leq \mu/(144T^2L^2)$, the preceding relation yields

$$
E \left[ \|x_{k+1} - x^*_k\|^2 \right] \leq \left( 1 - \frac{\mu}{9} - 8T^2L^2\alpha \right) T \alpha \alpha_k E \left[ \|x_k - x^*_k\|^2 \right]
$$

$$
+ 9\sigma^2T^2\alpha_k + \frac{3T^2\sigma^2f}{L|H|} \alpha_k
$$

$$
\leq \left( 1 - \frac{\mu T \alpha_k}{18} \right) E \left[ \|x_k - x^*_k\|^2 \right] + 9\sigma^2T^2\alpha_k + \frac{3T^2\sigma^2f}{L|H|} \alpha_k
$$

$$
\leq \left( 1 - \frac{\mu T \alpha_k}{18} \right)^k E \left[ \|x_0 - x^*_0\|^2 \right] + \frac{162\sigma^2T}{\mu} \alpha + \frac{54\sigma^2f}{\mu L|H|},
$$

which concludes our proof.
Byzantine fault-tolerance in federated local SGD under 2f-redundancy

C. Xie, O. Koyejo, and I. Gupta, “Fall of empires: Breaking Byzantine-tolerant SGD by inner product manipulation,” in Proc. 35th Uncertainty Artif. Intell. Conf., R. P. Adams and V. Gogate, Eds., vol. 115, 2020, pp. 261–270.

Y. Yin, Z. Chen, K. Ramchandran, and P. Bartlett, “Byzantine-robust distributed learning: Towards optimal statistical rates,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 5636–5645.

Y. Chen and E. J. Candès, “Solving random quadratic systems of equations is nearly as easy as solving linear systems,” Commun. Pure Appl. Math., vol. 70, no. 5, pp. 822–883, 2017.

H. Zhang, Y. Zhou, Y. Liang, and Y. Chi, “A nonconvex approach for phase retrieval: Reshaped Wirtinger flow and incremental algorithms,” J. Mach. Learn. Res., vol. 18, 2017.

R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” in Proc. Int. Conf. Mach. Learn., 2013, pp. 1310–1318.

R. Shokri and V. Shmatikov, “Privacy-preserving deep learning,” in Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur., 2015, pp. 1310–1321.

Z. Yang and W. U. Bajwa, “ByRDIE: Byzantine-resilient distributed coordinate descent for decentralized learning,” IEEE Trans. Signal Inf. Process. Netw., vol. 5, no. 4, pp. 611–627, Dec. 2019.

Y. Chen, L. Su, and J. Xu, “Distributed statistical machine learning in adversarial settings: Byzantine gradient descent,” in Proc. ACM Meas. Anal. Comput. Syst., 2017, vol. 1, no. 2, pp. 1–25.

S. Sundaram and B. Gharesifard, “Distributed optimization under adversarial nodes,” IEEE Trans. Autom. Control, vol. 64, no. 3, pp. 1063–1076, Mar. 2019.

C. Xie, O. Koyejo, and I. Gupta, “Phocas: Dimensional Byzantine-robust stochastic gradient descent,” CoRR, vol. abs/1805.09682, 2018. [Online]. Available: http://arxiv.org/abs/1805.09682

L. Li, W. Xu, T. Chen, G. B. Giannakis, and Q. Ling, “RSA: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets,” in Proc. AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 1544–1551.

J.-Y. Sohn, D.-J. Han, B. Choi, and J. Moon, “Election coding for distributed learning: Protecting SignsSGD against Byzantine attacks,” Adv. Neural Inf. Process. Syst., vol. 33, 2020.

I. Diakonikolas, G. Kamath, D. Kane, J. Li, J. Steinhardt, and A. Stewart, “Sever: A robust meta-algorithm for stochastic optimization,” in Proc. 36th Int. Conf. Mach. Learn., 2019, pp. 1596–1606.

A. Prasad, A. S. Suggala, S. Balakrishnan, and P. Ravikumar, “Robust estimation via robust gradient estimation,” J. Roy. Stat. Soc.: Ser. B. (Statistical Methodol.), vol. 82, no. 3, pp. 601–627, 2020.

L. Su and N. Vaidya, “Multi-agent optimization in the presence of byzantine adversaries: Fundamental limits,” in Proc. Amer. Control Conf., 2016, pp. 7183–7188.

M. Fang, X. Cao, J. Jia, and N. Gong, “Local model poisoning attacks to Byzantine-robust federated learning,” in Proc. 29th USENIX Secur. Symp., 2020, pp. 1605–1622.

L. Muñoz-González, K. T. Co, and E. C. Lupu, “Byzantine-robust federated machine learning through adaptive model averaging,” CoRR, vol. abs/1909.05125, 2019. [Online]. Available: http://arxiv.org/abs/1909.05125.

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.

C. Bajaj, “The algebraic degree of geometric optimization problems,” Discrete Comput. Geometry, vol. 5, no. 2, pp. 177–191, 1988.

M. B. Cohen, Y. T. Lee, G. Miller, J. Pachocki, and A. Sidford, “Geometric median in nearly linear time,” in Proc. 48th Ann. ACM Symp. Theory Comput., 2016, pp. 9–21.

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, Jul. 2020.

X. Cao, M. Fang, J. Liu, and N. Z. Gong, “Fltrust: Byzantine-robust federated learning via trust bootstrapping,” in Proc. ISOC Netw. Distrib. Syst. Secur. Symp., 2021, pp. 1–8.

S. Liu, N. Gupta, and N. H. Vaidya, “Approximate byzantine fault-tolerance in distributed optimization,” in Proc. ACM Symp. Princ. Distrib. Comput., 2021, pp. 379–389.

S. Farhadikhani, R. Guerraoui, N. Gupta, R. Pinot, and J. Stephan, “Byzantine machine learning made easy by resilient averaging of moments,” in Proc. 39th Int. Conf. Mach. Learn., 2022, pp. 6246–6283.

E. M. E. Mhamdi, R. Guerraoui, and S. Rouault, “Distributed momentum for Byzantine-resilient stochastic gradient descent,” in Proc. Int. Conf. Learn. Representations, 2021, pp. 1–20.

S. P. Karimireddy, L. He, and M. Jaggi, “Learning from history for Byzantine robust optimization,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 5311–5319.

N. Gupta and N. H. Vaidya, “Resilience in collaborative optimization: Redundant and independent cost functions,” CoRR, vol. abs/2003.09675, 2020. [Online]. Available: https://arxiv.org/abs/2003.09675

S. Liu, N. Gupta, and N. Vaidya, “Impact of redundancy on resilience in distributed optimization and learning,” in Proc. 24th Int. Conf. Distri. Comput. Netw., Khagpur, India, Jan. 4–7, 2023, pp. 80–89.

C. Xie, O. Koyejo, and I. Gupta, “Generalized Byzantine-tolerant SGD,” CoRR, vol. abs/1802.10116, 2018. [Online]. Available: http://arxiv.org/abs/1802.10116.

Nirupam Gupta received the bachelor’s degree in electrical engineering from the Indian Institute of Technology Kharagpur, Kharagpur, India, the master’s degree in mechanical engineering from the University of Maryland, College Park, MD, USA.

He is currently a Postdoctoral Fellow with the School of Computer Science, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland. He was a Postdoctoral Fellow of Computer Science with Georgetown University, Washington, DC, USA, from 2019 to 2021. His research interests include distributed machine learning: fault-tolerance and privacy.

Thinh T. Doan (Member, IEEE) received the bachelor’s degree in electrical engineering from the Hanoi University of Science and Technology, Hanoi, Vietnam, the master’s degree in electrical engineering from the University of Oklahoma, Norman, OK, USA, and the Ph.D. degree from the University of Illinois, Urbana-Champaign, Champaign, IL, USA.

He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, USA.

His research interests include intersection of control theory, optimization, machine learning, and applied probability theory with applications in robotics, wireless, and neuromorphic computing.

Nilin H. Vaidya (Fellow, IEEE) received the bachelor’s degree from the Birla Institute of Technology and Science-Pilani, Pilani, Rajasthan, India, the master’s degree from the Indian Institute of Science-Bangalore, Bengaluru, Karnataka, India, and the Ph.D. degree from the University of Massachusetts at Amherst, Amherst, MA, USA.

He is the Robert L. McDevitt, K.S.G., K.C.H.S., and Catherine H. McDevitt L.C.H.S., Chair of Computer Science with Georgetown University, Washington, DC, USA. He has held faculty positions with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Champaign, IL, USA, from 2001 to 2018, and with the Department of Computer Science, Texas A&M University, College Station, TX, USA, from 1992 to 2001. His research interests include distributed computing, distributed learning, and fault-tolerant algorithms.