2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Resilient Federated Learning under Byzantine Attack in Distributed Nonconvex Optimization with 2-f Redundancy

Amit Dutta Thinh T. Doan Jeffrey H. Reed

***Abstract*— We study the problem of Byzantine fault tolerance in a distributed optimization setting, where there is a group of N agents communicating with a trusted centralized coordinator. Among these agents, there is a subset of f agents that may not follow a prescribed algorithm and may share arbitrarily incorrect information with the coordinator. The goal is to find the optimizer of the aggregate cost functions of the honest agents. We will be interested in studying the local gradient descent method, also known as federated learning, to solve this problem. However, this method often returns an approximate value of the underlying optimal solution in the Byzantine setting. Recent work showed that by incorporating the so-called comparative elimination (CE) filter at the coordinator, one can provably mitigate the detrimental impact of Byzantine agents and precisely compute the true optimizer in the convex setting. The focus of the present work is to provide theoretical results to show the convergence of local gradient methods with the CE filter in a nonconvex setting. We will also provide a number of numerical simulations to support our theoretical results.**

I. Introduction

The constant expansion of large networks has resulted in an exponential growth of data, creating a pressing need for higher computational power and storage requirements. To address this demand, numerous distributed algorithms have been developed, where computation and data are distributed over networks. Federated learning has emerged as a popular distributed framework that facilitates collaborative training of a shared model by multiple devices [1], [2], [3]. This technique is particularly relevant in the context of wireless communication networks [4], [5], where there is a growing need for efficient and scalable machine learning solutions due to the rapidly increasing number of wireless devices. In this context, a common problem often reduces to optimize an aggregate objective function that is composed of N functions distributed at N different agents (e.g., mobile devices). By using federated learning framework, in applications like federated spectrum sensing over a wireless sensor network, the updates of any optimization algorithm can be implemented locally at the agents without requiring data being transmitted to the centralized coordinator (e.g., a server/network operator). By keeping data locally, federated learning not only reduces the amount of communications between agents and the server but also provides some level of privacy.

One of the main challenges in federated learning is the vulnerability of the system to malicious attacks where some agents in the network may fail or whose updates can be manipulated by an external entity. Such malicious (Byzantine) agents will have detrimental impacts to the performance of other agents, and if not addressed, it can lead to catastrophic failures of entire network [6].

In this paper, we study the performance of federated learning, in particular, the celebrated distributed local stochastic gradient descent (SGD), when a (small) number of agents in the network is malicious. We will focus on Byzantine malicious attacks, where Byzantine agents can observe the entire network and send any information to the centralized coordinator. Under the presence of Byzantine agents, it is impossible for nonfaulty agents to find the optimizer of the aggregate of functions at every agent (including Byzantine agents) as Byzantine agents can send a random number irrelevant to its function. Thus, we consider another meaningful objective in this setting, where the goal is to solve the optimization problem only involving the honest agents. In particular, we consider the setting where there are up to f faulty Byzantine agents with unknown identities. Our goal is address the following exact fault-tolerance problem.

Exact fault-tolerance problem: Let $\mathcal{H}$ be the set of honest agents with $|\mathcal{H}| \geq N - f$. A distributed optimization algorithm is said to have exact fault-tolerance if it allows all the non-faulty agents to compute

$$x^*_{\mathcal{H}} \in \operatorname*{argmin}_{x \in \mathbb{R}^d} \sum_{i \in \mathcal{H}} q^i(x). \tag{1}$$

We will study this exact fault-tolerance problem under the following $2f$-*redundancy* condition, which is necessary and sufficient for solving problem (1) [7].

Definition I.1 ($2f$-**redundancy**)**.** The set of non-faulty agents $\mathcal{H}$, with $|\mathcal{H}| \geq N - f$, is said to have $2f$-redundancy if for any subset $\mathcal{S} \subset \mathcal{H}$ with $|\mathcal{S}| \geq N - 2f$,

$$\operatorname*{argmin}_{x \in \mathbb{R}^d} \sum_{i \in \mathcal{S}} q^i(x) = \operatorname*{argmin}_{x \in \mathbb{R}^d} \sum_{i \in \mathcal{H}} q^i(x). \tag{2}$$

The definition above states that solving the optimization problem over the honest agents is equivalent to solving it over a minimum of $N - 2f$ honest agents. We will study the performance of distributed local SGD under the 2f-redundancy condition. Motivated by the recent work in [8], we will consider the so-called comparative elimination (CE) filter in the distributed local SGD to mitigate the detrimental impact of Byzantine agents. Our focus is to provide theoretical results to show the convergence of distributed local SGD with CE filter in a nonconvex setting.

The main contribution of this paper is to study the performance of federated local SGD with the CE filter for solving

The authors are with the Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA. Email: {amitdutta, thinhdoan, reedjh}@vt.edu

distributed nonconvex optimization problems under Byzantine attacks with the 2f-redundancy condition. We show that this method solves the exact fault-tolerance problem at a linear rate when the objective satisfies Polyak-Łojasiewicz (PL) condition. We will also provide a number of numerical simulations to illustrate our theoretical results.

A. Related work

Existing literature offers various Byzantine fault-tolerant aggregation schemes, such as *multi-KRUM* [9], *CWMT* [10], *GMoM* [11], *MDA* [12], and *Byzantine-RSA* [13] filters. However, these schemes don't guarantee exact fault-tolerance without additional assumptions. [8] showed the possibility of achieving exact fault tolerance in a deterministic setting and approximate fault tolerance in a stochastic setting with $2f$ redundancy. Recently, [14] proposed RESAM, a unified Byzantine fault-tolerant framework based on previous methods, demonstrating finite-time convergence with additional assumptions. Notably, their results apply to non-convex objectives but exclude the CE aggregation scheme.

Our work in this paper extends work by [8] to the non-convex setting, where the global objective function satisfies the PL condition. This broadens the analytical framework beyond strongly convex scenarios. Additionally, relevant work by [15], [16] addresses approximate fault tolerance with more relaxed conditions on Byzantine agents.

II. Federated Local SGD with Byzantine Agents

The proposed federated local SGD with CE filter is formally presented in Algorithm 1. Each user maintains a local variable $x^i_{k,t}$, estimating the local optimal solution at the k^{th} global iteration and t^{th} local iteration. The server maintains the global optimal solution estimate $\bar{x}_k$. The server initializes each user by transmitting an initial global model, $\bar{x}_0$. Each user i then runs $\mathcal{T}$ local SGD steps as shown in line 6, using a step-size α_k and a sample $g^i(x^i_{k,t})$ of its local gradient to update $x^i_{k,t}$. In the stochastic setting $g^i(x^i_{k,t}) = \nabla Q^i(x^i_{k,t}; \Delta^i_{k,t})$, where the gradients are i.i.d. sampled. After the local SGD steps, users transmit their estimates to the server. Using the CE filter the server removes all the values that are suspiciously large as compared to its value. Lines 10 and 11 show the CE filter where the server sorts the distance of the user estimates from its average and eliminates f largest distances. Finally, the server then computes a new average based on the estimates of remaining $N - f$ agents as shown in line 12.

Algorithm 1 Federated Local SGD with CE Filter

1: **Initialize**: The server initializes the model with $\bar{x}_0 \in \mathbb{R}^d$. Each agent initializes with step-sizes α_k and chooses $\mathcal{T}$.
2: **for** $k = 0, 1, ..$ **do**
3: All clients $i = 1, 2, .., N$ in parallel do
4: Receive $\bar{x}_k$ from the server and set $x^i_{k,0} = \bar{x}_k$
5: **for** $t = 0, ..., \mathcal{T} - 1$ **do**
6: $x^i_{k,t+1} = x^i_{k,t} - \alpha_k g^i(x^i_{k,t})$.
7: **end for**
8: Users send $x^i_{k,T}$ to the server
9: Server sorts these values as
10: $\|\bar{x}_k - x^{i_1}_{k,1}\| \leq \|\bar{x}_k - x^{i_2}_{k,1}\| \leq \cdots \leq \|\bar{x}_k - x^{i_N}_{k,1}\|$,
11: $\bar{x}_{k+1} = \frac{1}{|\mathcal{F}_k|} \sum_{i \in \mathcal{F}_k} x^i_{k,\mathcal{T}}$.
12: **end for**

III. Main Results

In this section we will provide our theoretical findings on the convergence properties of the local SGD with CE filter. For this we first define the average cost of the non-faulty agents defined as

$$q^{\mathcal{H}}(\bar{x}_k) = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} q^i(x), \tag{3}$$

Our main results are studied based on some fairly standard assumptions for non-convex optimization as stated below.

Assumption 1. (Lipchitz smoothness). For each $i \in \mathcal{H}$, q^i has L-Lipschitz continuous gradient

$$q^i(y) - q^i(x) \leq \nabla q^i(x)^T (y - x) + \frac{L}{2} \|y - x\|^2. \tag{4}$$

Assumption 2. The function $q^{\mathcal{H}}$ satisfies PL condition and quadratic growth with some $\mu \geq 0$

$$\|\nabla q^{\mathcal{H}}(\bar{x}_k)\|^2 \geq 2\mu(q^{\mathcal{H}}(\bar{x}_k) - q^{\mathcal{H}}(x^*_{\mathcal{H}})) \geq \mu^2 \|\bar{x}_k - x^*_{\mathcal{H}}\|^2. \tag{5}$$

Assumption 3. The The random variables Δ^i_k, for all i and k, are i.i.d., and there exists a positive constant σ such that

$$\mathbb{E}[\nabla Q^i(x, \Delta^i_{k,t}) | \mathcal{P}_{k,t}] = \nabla q^i(x), \quad \forall x \in \mathbb{R}^d,$$
$$\mathbb{E}[\|\nabla Q^i(x, \Delta^i_{k,t}) - \nabla q^i(x)\|^2 | \mathcal{P}_{k,t}] \leq \sigma^2, \quad \forall x \in \mathbb{R}^d. \tag{6}$$

In Assumption 3, $\mathcal{P}_{k,t}$ is defined as a filtration containing all the history generated by Algorithm 1 up to time $k + t$.

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

Further we have $|\mathcal{B}_k| + |\mathcal{H}_k| = |\mathcal{F}_k| = |\mathcal{H}|$ for any $k \geq 0$.

Next, we present our main theoretical result of this paper, where we study the convergence rate of Algorithm 1 in solving problem (1). For an ease of exposition, we present the proof of our result in Section VI.

From step 6 in Algorithm 1, for $i \in \mathcal{H}$ the local update is equivalent to

$$x^i_{k,t+1} = \bar{x}_k - \alpha_k \sum_{l=0}^{t} \nabla Q^i(x^i_{k,l}, \Delta^i_{k,l}). \tag{7}$$

Our result for the stochastic setting is presented below.

Theorem 1. Let $\{\bar{x}_k\}$ be generated by Algorithm 1. Let α_k be chosen as

$$\alpha_k = \alpha \leq \frac{\mu}{72L^2\mathcal{T}}. \tag{8}$$

Then, if the following condition holds

$$\frac{f}{N - f} \leq \frac{\mu}{3L}, \tag{9}$$

then we have

$$\mathbb{E}[q^{\mathcal{H}}(\bar{x}_{k+1}) - q^{\mathcal{H}}(x^*_{\mathcal{H}})] \leq \left(1 - \frac{\alpha_k \mu \mathcal{T}}{36}\right)^{k+1} \mathbb{E}[q^{\mathcal{H}}(\bar{x}_0) - q^{\mathcal{H}}(x^*_{\mathcal{H}})] + \frac{180 L \mathcal{T} \alpha_k \sigma^2}{\mu} + \frac{72 \mathcal{T} \sigma^2 f}{\mu |\mathcal{H}|}. \quad (10)$$

Remark 1. In Theorem 1 due to the constant step size, the optimality error converges linearly only to a ball centered at the origin. The size of the ball is determined by two factors. The first factor is dependent on the step size α, which is commonly observed in the convergence of local gradient descent with non-faulty agents. The second factor is influenced by the level of gradient noise, denoted by σ. This noise is a result of both the Byzantine agents and the stochastic gradient samples.

It is worth noting that our comparative filter is specifically designed to eliminate potentially erroneous values sent by the Byzantine agents. However, it is unable to address the issue of variance in their stochastic samples. One possible solution to this problem is to have each agent sample a mini-batch of size m, thereby replacing σ^2 in (10) with σ^2/m. By increasing the size of m, the optimality error can be made arbitrarily close to zero. Furthermore, when $\alpha_k \sim 1/k$, the convergence rate is $\mathcal{O}(1/k)$.

Finally, if we can have access to the exact values of the gradients ∇q^i, then $\bar{x}_k$ converges exactly to $x^\star$ exponentially.

IV. Simulations

For the evaluation of the federated local SGD with CE filter, we consider a network with $N = 50$ agents with a varying number of byzantine agents. We further compare the performance of the convergence of the algorithm with various other existing byzantine filters, namely multi-KRUM [9] and Coordinate-Wise Trimmed Median (CWTM) [17], [10]. Our experiment goals are the following:

- Fix the number of byzantine agents compare the performace of the algorithm with differnt byzantine filters for $\mathcal{T} = 3$ local local and 50 global communication rounds.
- For $\mathcal{T} = 3$ local local and 50 global communication rounds compare the performance of local SGD with CE filter with varying number of byzantine agents, $f = 2, 5, 8, 10$.

We consider a scenario where we have 50 agents trying to optimize sum of given local functions. We present results from Algorithm 1 and local SGD with the aforementioned byzantine filters. Here at each local iteration any agent i has access to an i.i.d sample of it's local gradients. First, we consider a regression problem

$$\min_x \sum_{i=1}^{N} q^i(x) \triangleq \sum_{i=1}^{N} \left(\|\mathbf{A}^i x - b^i\|^2 + \sin^2(\|\mathbf{A}^i x - b^i\|) \right). \quad (11)$$

where $\mathbf{A}_i$ and b_i are the feature vector and labels respectively for i^{th} agent. This is an example of an invex but non-convex function satisfying the PL condition. Here each agent has its own estimate $x_i \in \mathbb{R}^d$ and also maintains the parameters $(\mathbf{A}_i, b_i)$ with $\mathbf{A}_i \in \mathbb{R}^{l \times d}$ and $b_i \in \mathbb{R}^l$. Further, each agent has an associated local function $q_i(x)$ where x is the decision variable, $\mathbf{A}_i$ is the weight or importance of agent i's objective, and b_i is the target or reference value of user i's objective. Here we further note that the since objectives satisfies the PL condition, the optimal solution x^* will be unique for any set of honest agent $\mathcal{H}$ (see Remark 1 in [8]).

Second, we consider

$$\min_x \sum_{i=1}^{N} q_i(x) = \min_x \sum_{i=1}^{N} \left(\frac{1}{1 + \exp\left(-\|\mathbf{A}_i x - b_i\|\right)} \right), \quad (12)$$

where the objectives are non-convex and do not satisfy the PL condition. Here we observe variation of the term $\frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \|\nabla q^i(\bar{x}_k)\|^2$. This is a standard error term used for study analysis of a non-convex optimization problem. The simulation results are shown in Fig. 1 and 2 respectively. Fig.1 show the performances of CWTM, Multi-KRUM and CE filer with non-faulty Local GD as baseline for comparison as we vary the number of byzantine agents $f = 2, 5, 8, 10$. We observe that Local SGD with CE filter out performs other byzantine filters. We further note that as the byzantine agent increase the convergence error increases. Fig 2 and 3 shows the performance of CE filter for local iterations $\mathcal{T} = 1, 3$. We conclude that as we increase the local iterations, Algorithm 1 converges faster which is consistent with previous work [8]. The observed results are in line with our theoretical findings, which demonstrate that achieving only an approximate fault tolerance in stochastic scenarios is possible.

V. Acknowledgement

The work of Amit Dutta and Thinh T. Doan was partially supported by the Commonwealth Cyber Initiative, an investment in the advancement of cyber R&D, innovation, and workforce development. For more information about CCI, visit www.cyberinitiative.org.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[2] H. Zhao, Z. Li, and P. Richtárik, "Fedpage: A fast local stochastic gradient method for communication-efficient federated learning," *arXiv preprint arXiv:2108.04755*, 2021.

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

[4] S. Niknam, H. S. Dhillon, and J. H. Reed, "Federated learning for wireless communications: Motivation, opportunities, and challenges," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 46–51, 2020.

[5] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 1935–1949, 2020.

[6] S. Shi, Y. Xiao, W. Lou, C. Wang, X. Li, Y. T. Hou, and J. H. Reed, "Challenges and new directions in securing spectrum access systems," *IEEE Internet of Things Journal*, vol. 8, no. 8, pp. 6498–6518, 2021.

[7] N. Gupta and N. H. Vaidya, "Fault-tolerance in distributed optimization: The case of redundancy," in *Proceedings of the 39th Symposium on Principles of Distributed Computing*, 2020, pp. 365–374.

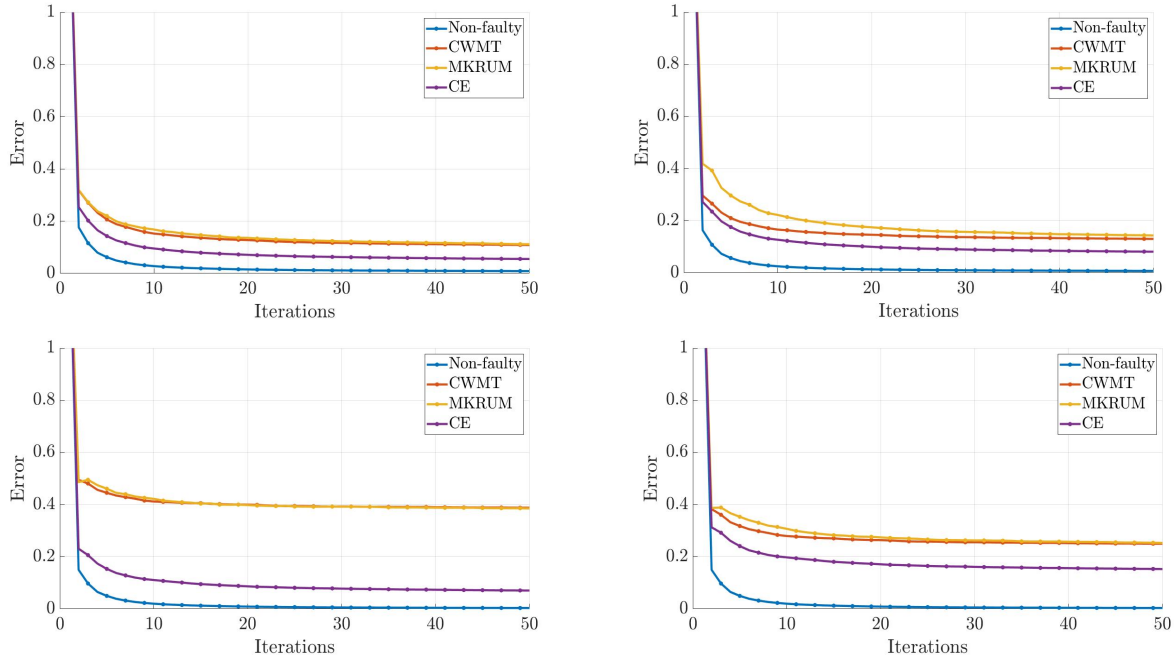

Fig. 1: The plots show the error $q^{\mathcal{H}}(\bar{x}_k) - q^{\mathcal{H}}(x^*)$ for $\mathcal{T} = 3$ for local iterations of local SGD with CE filter (Algorithm 1), CWMT, and Multi-KRUM for problem (11). Going clockwise we vary the byzantine agents as $f = 2, 5, 8, 10$.

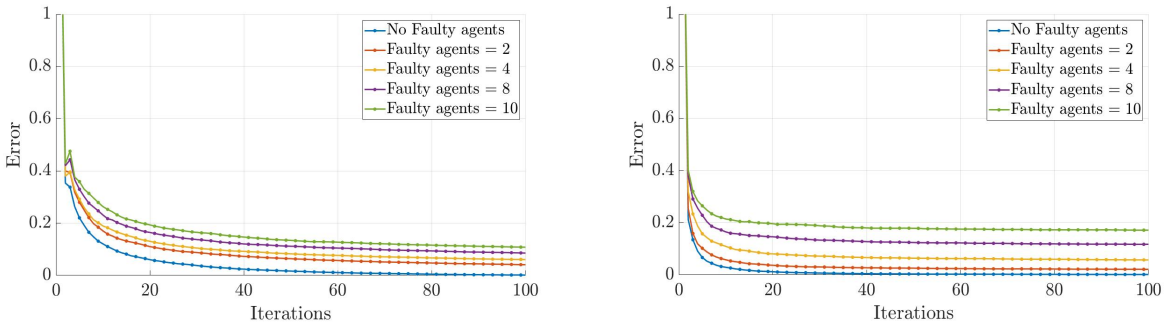

Fig. 2: The plots show the error $q^{\mathcal{H}}(\bar{x}_k) - q^{\mathcal{H}}(x^*)$ for problem (11) using Algorithm 1. Left and right figures show simulations with the local iterations $\mathcal{T} = 1$ and $\mathcal{T} = 3$, respectively.

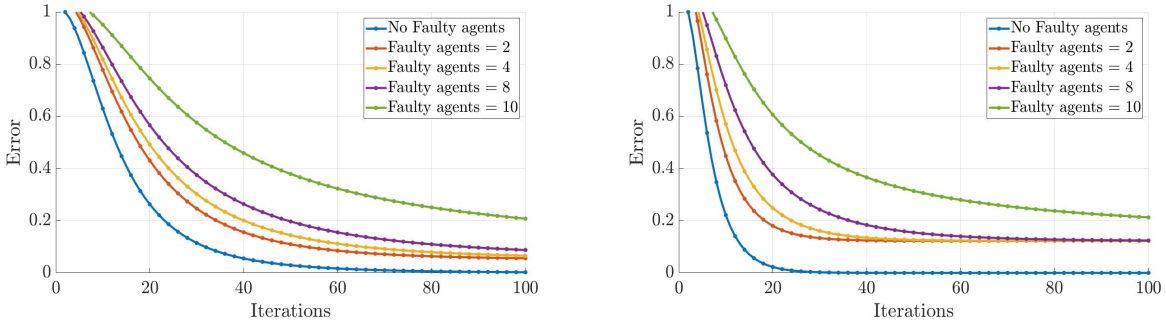

Fig. 3: The plots show the error $\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \|\nabla q^i(\bar{x}_k)\|^2$ for local SGD with CE filter in solving problem (12). Left and right figures show simulations with the local iterations $\mathcal{T} = 1$ and $\mathcal{T} = 3$, respectively.

[8] N. Gupta, T. T. Doan, and N. Vaidya, "Byzantine fault-tolerance in federated local sgd under $2f$-redundancy," *IEEE Transactions on Control of Network Systems*, 2023.
[9] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," *Advances in neural information processing systems*, vol. 30, 2017.
[10] L. Su and S. Shahrampour, "Finite-time guarantees for byzantine-resilient distributed state estimation with noisy measurements," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3758–3771, 2019.
[11] Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 1, no. 2, pp. 1–25, 2017.
[12] R. Guerraoui, S. Rouault, *et al.*, "The hidden vulnerability of distributed learning in byzantium," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3521–3530.
[13] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 1544–1551.
[14] S. Farhadkhani, R. Guerraoui, N. Gupta, R. Pinot, and J. Stephan, "Byzantine machine learning made easy by resilient averaging of momentums," in *International Conference on Machine Learning*. PMLR, 2022, pp. 6246–6283.
[15] X. Cao and L. Lai, "Distributed gradient descent algorithm robust to an arbitrary number of byzantine attackers," *IEEE Transactions on Signal Processing*, vol. 67, no. 22, pp. 5850–5864, 2019.
[16] ——, "Distributed approximate newton's method robust to byzantine attackers," *IEEE Transactions on Signal Processing*, vol. 68, pp. 6011–6025, 2020.
[17] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5650–5659.

VI. APPENDIX

A. Proof of Theorem 1

Proof. Using the local SGD update for $i \in \mathcal{H}$ and Assumptions 1–(3) we obtain the relations below. We skip their proofs due to space limitations.

$$\begin{aligned}
&\mathbb{E}[\|x^i_{k,t+1} - x^*_{\mathcal{H}}\|] \le \frac{2}{\mu}\mathbb{E}[\|\nabla q^{\mathcal{H}}(\bar{x}_k)\|] + \frac{2\sigma}{L},\\
&\mathbb{E}[\|x^i_{k,t+1} - \bar{x}_k\|] \le \frac{2L\mathcal{T}\alpha_k}{\mu}\mathbb{E}[\|\nabla q^{\mathcal{H}}(\bar{x}_k)\|] + 3\sigma\mathcal{T}\alpha_k,\\
&\mathbb{E}\big[\|x^i_{k,t+1} - x^*_{\mathcal{H}}\|^2\big] \le \frac{2}{\mu^2}\mathbb{E}[\|\nabla q^{\mathcal{H}}(\bar{x}_k)\|^2] + \frac{2\sigma^2}{L^2}.
\end{aligned} \tag{13}$$

By (4), we have

$$\begin{aligned}
\mathbb{E}[q^{\mathcal{H}}(\bar{x}_{k+1}) - q^{\mathcal{H}}(\bar{x}_k)] \le \mathbb{E}[\nabla q^{\mathcal{H}}(\bar{x}_k)^T(\bar{x}_{k+1} - \bar{x}_k)]\\
+ \frac{L}{2}\mathbb{E}[\|\bar{x}_{k+1} - \bar{x}_k\|^2].
\end{aligned} \tag{14}$$

We next analyze each term on the right-hand side of (14). Note that $|\mathcal{F}_k| = |\mathcal{H}| - N - f$ and $|\mathcal{B}_k| = |\mathcal{H}\backslash\mathcal{H}_k|$. Thus, we have the following relation

$$\begin{aligned}
\bar{x}_{k+1} &= \frac{1}{|\mathcal{H}|}\Big[\sum_{i\in\mathcal{H}} x^i_{k,\mathcal{T}} + \sum_{i\in\mathcal{B}_k} x^i_{k,\mathcal{T}} - \sum_{i\in\mathcal{H}\backslash\mathcal{H}_k} x^i_{k,\mathcal{T}}\Big]\\
&\overset{(7)}{=} \bar{x}_k - \frac{\alpha_k}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{t=0}^{\mathcal{T}-1}\nabla Q^i(x^i_{k,t};\Delta^i_{k,t}) + V_x\\
&= \bar{x}_k - \mathcal{T}\alpha_k\nabla q^{\mathcal{H}}(\bar{x}_k) + V_x\\
&\quad - \frac{\alpha_k}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{t=0}^{\mathcal{T}-1}(\nabla Q^i(x^i_{k,t};\Delta^i_{k,t}) - \nabla q^i(\bar{x}_k)).
\end{aligned}$$

where

$$V_x = \frac{1}{|\mathcal{H}|}\Big[\sum_{i\in\mathcal{B}_k}(x^i_{k,\mathcal{T}} - \bar{x}_k) - \sum_{i\in\mathcal{H}\backslash\mathcal{H}_k}(x^i_{k,\mathcal{T}} - \bar{x}_k)\Big]. \tag{15}$$

We next consider the first term in (14)

$$\begin{aligned}
&\nabla q^{\mathcal{H}}(\bar{x}_k)^T(\bar{x}_{k+1} - \bar{x}_k)\\
&\le -\frac{\alpha_k}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{t=0}^{\mathcal{T}-1}\nabla q^{\mathcal{H}}(\bar{x}_k)^T(\nabla Q^i(x^i_{k,t};\Delta^i_{k,t}) - \nabla q^i(\bar{x}_k))\\
&\quad - \mathcal{T}\alpha_k\|\nabla q^{\mathcal{H}}(\bar{x}_k)\|^2 + \nabla q^{\mathcal{H}}(\bar{x}_k)^T V_x.
\end{aligned} \tag{16}$$

Analyzing the first term on the right-hand side of the above inequality we have,

$$\begin{aligned}
&\frac{-\alpha_k}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{t=0}^{\mathcal{T}-1}\mathbb{E}[\nabla q^{\mathcal{H}}(\bar{x}_k)^T(\nabla Q^i(x^i_{k,t};\Delta^i_{k,t})\\
&- \nabla q^i(\bar{x}_k))|\mathcal{P}_{k,t}]\\
&= -\frac{\alpha_k}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{t=0}^{\mathcal{T}-1}\nabla q^{\mathcal{H}}(\bar{x}_k)^T(\nabla q^i(x^i_{k,t}) - \nabla q^i(\bar{x}_k))\\
&\le \frac{L\alpha_k^2}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{t=0}^{\mathcal{T}-1}\|\bar{x}_k - x^*_{\mathcal{H}}\|\|x^i_{k,t} - \bar{x}_k\|\\
&\le \frac{L^3\alpha_k^2\mathcal{T}^2}{2\mu^2}\|\nabla q^{\mathcal{H}}(\bar{x}_k)\|^2 + \frac{L}{2\mathcal{T}|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{t=0}^{\mathcal{T}-1}\|x^i_{k,t} - \bar{x}_k\|^2,
\end{aligned}$$

where the last inequality is due to Cauchy-Schwarz inequality $2xy \le \eta x^2 + y^2/\eta$ for any $\eta > 0$ and $x, y \in \mathbb{R}$. Taking an expectation on both sides of the above inequality and using (13) we have

$$\begin{aligned}
&-\frac{\alpha_k}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{t=0}^{\mathcal{T}-1}\mathbb{E}[\nabla q^{\mathcal{H}}(\bar{x}_k)^T(\nabla Q^i(x^i_{k,t};\Delta^i_{k,t}) - \nabla q^i(\bar{x}_k))]\\
&\le \frac{3L^3\alpha_k^2\mathcal{T}^2}{2\mu^2}\mathbb{E}[\|\nabla q^{\mathcal{H}}(\bar{x}_k)\|^2] + \frac{3L\mathcal{T}^2\alpha_k^2\sigma^2}{2}.
\end{aligned}$$

Next, we analyze $\|V_x\|^2$. Using (15) and the fact that there exists an agent $j \in \mathcal{H}\backslash\mathcal{H}_k$ such that $\|x^i_{k,t} - \bar{x}_k\| \le \|x^j_{k,t} - \bar{x}_k\|$ for all agent $i \in \mathcal{B}_k$, we obtain

$$\begin{aligned}
&\mathbb{E}\|V_x\|^2\\
&= \mathbb{E}\Big[\Big\|\frac{1}{|\mathcal{H}|}\Big(\sum_{i\in\mathcal{B}_k}(x^i_{k,\mathcal{T}} - \bar{x}_k) - \sum_{i\in\mathcal{H}\backslash\mathcal{H}_k}(x^i_{k,\mathcal{T}} - \bar{x}_k)\Big)\Big\|^2\Big]\\
&\le \frac{2|\mathcal{B}_k|^2}{|\mathcal{H}|^2}\mathbb{E}[\|x^j_{k,t} - \bar{x}_k\|^2] + \frac{2|\mathcal{B}_k|}{|\mathcal{H}|^2}\sum_{i\in\mathcal{H}\backslash\mathcal{H}_k}\mathbb{E}[\|x^i_{k,t} - \bar{x}_k\|^2],\\
&\overset{(13)}{\le} \frac{8L^2\mathcal{T}^2\alpha_k^2}{\mu^2}\frac{|\mathcal{B}_k|^2}{|\mathcal{H}|^2}\mathbb{E}[\|\nabla q^{\mathcal{H}}(\bar{x}_k)\|^2] + \frac{12\mathcal{T}^2\alpha_k^2\sigma^2|\mathcal{B}_k|^2}{|\mathcal{H}|^2}.
\end{aligned}$$

We now analyze the last term in the right-hand side of (16). Using the above result and the relation $< x, y > \eta\|x\|^2/2 + \|y\|^2/2\eta \le$ for any $\eta > 0$ we have

$$\begin{aligned}
&\mathbb{E}[\nabla q^{\mathcal{H}}(\bar{x}_k)^T V_x]\\
&\le \frac{3L\mathcal{T}\alpha_k|\mathcal{B}_k|}{2\mu|\mathcal{H}|}\mathbb{E}\big[\|\nabla q^{\mathcal{H}}(\bar{x}_k)\|^2\big] + \frac{\mu|\mathcal{H}|}{6L\mathcal{T}\alpha_k|\mathcal{B}_k|}\mathbb{E}\big[\|V_x\|^2\big],\\
&\le \frac{17L\mathcal{T}\alpha_k|\mathcal{B}_k|}{6\mu|\mathcal{H}|}\mathbb{E}[\|\nabla q^{\mathcal{H}}(\bar{x}_k)\|^2] + \frac{2\mathcal{T}\mu\alpha_k\sigma^2|\mathcal{B}_k|}{L|\mathcal{H}|}.
\end{aligned}$$

Thus we obtain from (16)

$$\mathbb{E}[\nabla q^{\mathcal{H}}(\bar{x}_{k+1})^T(\bar{x}_{k+1}-\bar{x}_k)]$$
$$\leq \Big(-\alpha_k\mathcal{T}+\frac{17L\mathcal{T}\alpha_k|\mathcal{B}_k|}{6\mu|\mathcal{H}|}+\frac{3L^3\alpha_k^2\mathcal{T}^2}{2\mu^2}\Big)\mathbb{E}[\|\nabla q^{\mathcal{H}}(\bar{x}_k)\|^2]$$
$$+\frac{3L\mathcal{T}^2\alpha_k^2\sigma^2}{2}+\frac{2\mathcal{T}\mu\alpha_k\sigma^2|\mathcal{B}_k|}{L|\mathcal{H}|}. \tag{17}$$

Next we analyze the second term on right hand side of (14)

$$\|\bar{x}_{k+1}-\bar{x}_k\|^2$$
$$=\Big\|\frac{\alpha_k}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{t=0}^{\mathcal{T}-1}\nabla Q^i(x_{k,t}^i;\Delta_{k,t}^i)\Big\|^2+\|V_x\|^2$$
$$-\frac{2\alpha_k}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{t=0}^{\mathcal{T}-1}V_x^T\nabla Q^i(x_{k,t}^i;\Delta_{k,t}^i). \tag{18}$$

Taking the expectation of the first term on the right-hand side of (18)

$$\mathbb{E}\Big[\Big\|\frac{\alpha_k}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{t=0}^{\mathcal{T}-1}\nabla Q^i(x_{k,t}^i;\Delta_{k,t}^i)\Big\|^2\Big]$$
$$\leq\frac{\alpha_k^2\mathcal{T}}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{t=0}^{\mathcal{T}-1}\mathbb{E}\Big[\Big\|\nabla Q^i(x_{k,t}^i;\Delta_{k,t}^i)\Big\|^2\Big]$$
$$\leq\frac{\alpha_k^2\mathcal{T}}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{t=0}^{\mathcal{T}-1}\mathbb{E}\Big[\Big\|\nabla Q^i(x_{k,t}^i;\Delta_{k,t}^i)-\nabla q^i(x_{k,t}^i)\Big\|^2\Big]$$
$$+\frac{\alpha_k^2\mathcal{T}}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{t=0}^{\mathcal{T}-1}\mathbb{E}\Big[\Big\|\nabla q^i(x_{k,t}^i)-\nabla q^i(x_{\mathcal{H}}^*)\Big\|^2\Big]$$
$$\leq\frac{L^2\mathcal{T}\alpha_k^2}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{t=0}^{\mathcal{T}-1}\mathbb{E}\Big[\Big\|x_{k,t}^i-x_{\mathcal{H}}^*\Big\|^2\Big]+\mathcal{T}^2\sigma^2\alpha_k^2$$
$$\overset{(13)}{\leq}\frac{2L^2\mathcal{T}^2\alpha_k^2}{\mu^2}\mathbb{E}\Big[\Big\|\nabla q^{\mathcal{H}}(\bar{x}_k)\Big\|^2\Big]+3\mathcal{T}^2\sigma^2\alpha_k^2. \tag{19}$$

Now we analyze the last term in the right-hand side of (18). For this using the relation $<x,y>\eta\|x\|^2/2+\|y\|^2/2\eta\leq$ for any $\eta>0$ we have

$$\frac{-2}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{t=0}^{\mathcal{T}-1}\mathbb{E}[\alpha_k V_x^T\nabla Q^i(x_{k,t}^i;\Delta_{k,t}^i)]$$
$$\leq\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{t=0}^{\mathcal{T}-1}\Big(\frac{1}{\mathcal{T}}\mathbb{E}[\|V_x\|^2]+\mathcal{T}\alpha_k^2\mathbb{E}[\|\nabla Q^i(x_{k,t}^i;\Delta_{k,t}^i)\|^2]\Big)$$
$$\leq\frac{\mathcal{T}\alpha_k^2}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{t=0}^{\mathcal{T}-1}\mathbb{E}[\|\nabla Q^i(x_{k,t}^i;\Delta_{k,t}^i)-\nabla q^i(x_{k,t}^i)\|^2]$$
$$+\frac{\mathcal{T}\alpha_k^2}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{t=0}^{\mathcal{T}-1}\mathbb{E}[\|\nabla q^i(x_{k,t}^i)-\nabla q^i(x_{\mathcal{H}}^*)\|^2]+\mathbb{E}[\|V_x\|^2]$$
$$\leq\mathbb{E}\|V_x\|^2+\frac{\mathcal{T}L^2\alpha_k^2}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\sum_{t=0}^{\mathcal{T}-1}\mathbb{E}[\|x_{k,t}^i-x_{\mathcal{H}}^*\|^2]+\mathcal{T}^2\sigma^2\alpha_k^2$$
$$\overset{(13)}{\leq}\mathbb{E}\big[\|V_x\|^2\big]+\frac{2L^2\mathcal{T}^2\alpha_k^2}{\mu^2}\mathbb{E}\big[\|\nabla q^{\mathcal{H}}(\bar{x}_k)\|^2\big]+3\mathcal{T}^2\sigma^2\alpha_k^2. \tag{20}$$

Substituting (19) and (20) into (18) we get obtain

$$\frac{L}{2}\mathbb{E}[\|\bar{x}_{k+1}-\bar{x}_k\|^2]$$
$$\leq 2\mathbb{E}\big[\|V_x\|^2\big]+\frac{4L^2\mathcal{T}^2\alpha_k^2}{\mu^2}\mathbb{E}\big[\|\nabla q^{\mathcal{H}}(\bar{x}_k)\|^2\big]+6\mathcal{T}^2\sigma^2\alpha_k^2$$
$$\leq\Big(\frac{2L^3\mathcal{T}^2\alpha_k^2}{\mu^2}+\frac{4L^3\mathcal{T}^2\alpha_k^2}{\mu^2}\frac{|\mathcal{B}_k|^2}{|\mathcal{H}|^2}\Big)\mathbb{E}[\|\nabla q^{\mathcal{H}}(\bar{x}_k)\|^2]$$
$$+3L\mathcal{T}^2\sigma^2\alpha_k^2+\frac{6L\mathcal{T}^2\alpha_k^2\sigma^2|\mathcal{B}_k|^2}{|\mathcal{H}|^2}. \tag{21}$$

Finally, substituting (17) and (21) into (14) and using $|\mathcal{B}_k|\leq f$ we get

$$\mathbb{E}[q^{\mathcal{H}}(\bar{x}_{k+1})-q^{\mathcal{H}}(\bar{x}_k)]$$
$$\leq\Big(-\Big(1-\frac{17Lf}{6\mu|\mathcal{H}|}\Big)\alpha_k\mathcal{T}+\frac{7L^3\mathcal{T}^2\alpha_k^2}{2\mu^2}$$
$$+\frac{4L^3\mathcal{T}^2\alpha_k^2f^2}{\mu^2|\mathcal{H}|^2}\Big)\mathbb{E}\|\nabla q^{\mathcal{H}}(\bar{x}_k)\|^2+\frac{9L\mathcal{T}^2\sigma^2\alpha_k^2}{2}$$
$$+\frac{2\mathcal{T}^2\mu\alpha_k\sigma^2f}{L|\mathcal{H}|}+\frac{6L\mathcal{T}^2\sigma^2\alpha_k^2f^2}{|\mathcal{H}|^2}.$$

Since $\frac{f}{|\mathcal{H}|}\leq\frac{\mu}{3L}$, we obtain (9)

$$1-\frac{17Lf}{6|\mathcal{H}|}>\frac{1}{18},$$

using which we have

$$\mathbb{E}[q^{\mathcal{H}}(\bar{x}_{k+1})-q^{\mathcal{H}}(\bar{x}_k)]$$
$$\leq\Big(\frac{-\alpha_k\mathcal{T}}{18}+\Big(\frac{7}{2}+\frac{4f^2}{|\mathcal{H}|^2}\Big)\frac{L^2\mathcal{T}^2\alpha_k^2}{\mu^2}\Big)\mathbb{E}\big[\|\nabla q^{\mathcal{H}}(\bar{x}_k)\|^2\big]$$
$$+5L\mathcal{T}^2\sigma^2\alpha_k^2+\frac{2\mathcal{T}^2\alpha_k\mu\sigma^2f}{L|\mathcal{H}|}.$$

Next, using $\alpha_k=\alpha\leq\frac{\mu^2}{72L^3\mathcal{T}}$ from (8) we have,

$$\mathbb{E}[q^{\mathcal{H}}(\bar{x}_{k+1})-q^{\mathcal{H}}(\bar{x}_k)]$$
$$\leq-\frac{\alpha_k\mathcal{T}}{36}\mathbb{E}\big[\|\nabla q^{\mathcal{H}}(\bar{x}_k)\|^2\big]+5L\mathcal{T}^2\sigma^2\alpha_k^2+\frac{2\mathcal{T}^2\alpha_k\mu\sigma^2f}{L|\mathcal{H}|},$$

which gives us

$$\mathbb{E}[q^{\mathcal{H}}(\bar{x}_{k+1})-q^{\mathcal{H}}(x_{\mathcal{H}}^*)]$$
$$\leq\Big(1-\frac{\alpha_k\mu\mathcal{T}}{36}\Big)\mathbb{E}[q^{\mathcal{H}}(\bar{x}_k)-q^{\mathcal{H}}(x_{\mathcal{H}}^*)]$$
$$+5L\mathcal{T}^2\sigma^2\alpha_k^2+\frac{2\mathcal{T}^2\alpha_k\mu\sigma^2f}{L|\mathcal{H}|},$$
$$\leq\Big(1-\frac{\alpha_k\mu\mathcal{T}}{36}\Big)^{k+1}\mathbb{E}[q^{\mathcal{H}}(\bar{x}_0)-q^{\mathcal{H}}(x_{\mathcal{H}}^*)]$$
$$+\frac{180L\mathcal{T}\alpha_k\sigma^2}{\mu}+\frac{72\mathcal{T}\sigma^2f}{\mu|\mathcal{H}|}. \tag{22}$$

This concludes our proof. □