BYZANTINE-ROBUST AGGREGATION WITH GRADIENT DIFFERENCE COMPRESSION AND STOCHASTIC VARIANCE REDUCTION FOR FEDERATED LEARNING

Heng Zhu[1,2] *Qing Ling*[1]

[1]Sun Yat-Sen University [2]University of California, San Diego

ABSTRACT

We investigate the problem of Byzantine-robust compressed federated learning, where the transmissions from the workers to the master node are compressed, and subject to malicious attacks from an unknown number of Byzantine workers. We show that the vanilla combination of the distributed compressed stochastic gradient descent (SGD) with geometric median-based robust aggregation suffers from the compression noise under Byzantine attacks. In light of this observation, we propose to reduce the compression noise with gradient difference compression to improve the Byzantine-robustness. We also observe the impact of the intrinsic stochastic noise from selecting random samples, and adopt the stochastic average gradient algorithm (SAGA) to gradually eliminate the inner variations of regular workers. We prove that the proposed algorithm reaches a neighborhood of the optimal solution at a linear convergence rate, and the asymptotic learning error is in the same order as that of the state-of-the-art uncompressed method. Finally, numerical experiments demonstrate the effectiveness of the proposed method.

***Index Terms*—** Federated learning, communication efficiency, Byzantine robustness, gradient compression.

1. INTRODUCTION

With the rapid development of intelligent devices, federated learning has been proposed as an effective approach to fusing local data of distributed devices without jeopardizing data privacy. In a federated learning system, local data are kept at the distributed devices (also termed as workers). At each iteration, the workers send local stochastic gradients to a master node; the latter aggregates the local stochastic gradients to update the trained model [1, 2, 3, 4]. Beyond data privacy, communication efficiency and robustness to various adversarial attacks are also major concerns of federated learning.

Information exchange between the workers and the master node, especially transmitting the local stochastic gradients from the workers to the master node, is a bottleneck of a federated learning system. In particular, when the trained model is high-dimensional, the local stochastic gradients are high-dimensional too and the communication burden is remarkable. To improve the communication efficiency, several popular strategies have been proposed. One such strategy is to reduce the communication frequency by performing multiple rounds of local updates before one round of transmissions [5, 6, 7]. Another orthogonal strategy is to reduce the sizes of transmitted messages by compression. Typical compression methods include quantization that uses limited bits to represent real vectors [8], and sparsification that enforces sparsity of transmitted vectors [9]. In this work we focus on compression. At each iteration, the workers compress the local stochastic gradients and send to the master node. Then, the master node aggregates the received compressed local stochastic gradients to obtain a new direction.

In a federated learning system, however, transmitting the compressed local stochastic gradients is vulnerable to adversarial attacks [10, 11, 12, 13]. Not all the workers are guaranteed to be reliable and send the true compressed local stochastic gradients. Some of them may send faulty messages to bias the aggregation and lead the optimization process to a wrong direction. To characterize the attacks, we consider the Byzantine attacks model where the number and identities of Byzantine workers are unknown to the master node. The Byzantine workers are assumed to be omniscient, can collude with each other, and may send arbitrary malicious messages [14]. To defend against Byzantine attacks, several robust aggregation rules have been proposed to replace the mean aggregation rule in the popular distributed stochastic gradient descent (SGD) algorithm [15, 16, 17]. These approaches are provably able to alleviate the influence of the malicious messages sent by the Byzantine workers on the optimization process.

In this paper, we investigate the problem of Byzantine-robust federated learning with compression, simultaneously considering Byzantine-robustness and communication efficiency. For Byzantine-robust aggregation rules, the noise from compressing the local stochastic gradients significantly weakens their ability to defend against Byzantine attacks. To theoretically justify this claim, we show that even with unbiased compressors, the naive combination of the distributed compressed stochastic gradient descent (SGD) with geometric median-based robust aggregation still suffers from the compression noise. This observation illustrates the necessity of

Qing Ling (corresponding author) is supported by NSF China Grant 61973324, Guangdong Basic and Applied Basic Research Foundation Grant 2021B1515020094, and Guangdong Province Key Laboratory of Computational Science Grant 2020B1212060032.

reducing the compression noise under Byzantine attacks. In addition, the stochastic noise caused by selecting random samples to compute the local stochastic gradients also brings difficulties to handling Byzantine attacks [18, 19, 20]. To address these two issues, we propose a novel algorithm, termed as Byzantine-RObust Aggregation with gradient Difference Compression And STochastic variance reduction (BROADCAST), to reduce both compression and stochastic noise. To be specific, we apply gradient difference compression [21, 22] to reduce the compression noise, and adopt the stochastic average gradient algorithm (SAGA) [23] to gradually eliminate the inner variations of regular workers.

For existing Byzantine-robust methods with compression, [24] shows that SignSGD is able to handle a certain class of Byzantine attacks. However, we will show in the numerical experiments that it fails upon several common Byzantine attacks. In [25], gradient norm thresholding is used to remove potential malicious messages with compression, where error feedback is applied to reduce the learning error and Gaussian attacks are tested. However, gradient norm thresholding removes a given fraction of messages. In contrast, our proposed algorithm does not need any prior knowledge about the number of Byzantine workers. In addition, gradient norm thresholding can be viewed as a modified mean aggregation rule. Analyzing its error feedback extension is rather straightforward, and relies on the assumption of bounded stochastic gradients. Our analysis considers the combination of geometric median and gradient difference compression, and is hence more challenging. Further, we do not require the assumption of bounded stochastic gradients.

Due to the page limit, proofs and additional experiments are delegated to an extended version of this paper [26].

2. PROBLEM FORMULATION

Consider a distributed federated learning system with one master node and W workers in a set $\mathcal{W}$. Among these workers, R of them are regular and constitute a set $\mathcal{R}$, while the rest B of them are Byzantine and constitute a set $\mathcal{B}$. Note that the identities of regular and Byzantine workers are unknown. The Byzantine workers are assumed to be omniscient and can collude with each other to send arbitrary malicious messages to the master node. The problem of interest is to find an optimal solution to the finite-sum optimization problem

$$x^* = \arg\min_x f(x) := \frac{1}{R}\sum_{\omega\in\mathcal{R}} f_\omega(x), \tag{1}$$

with $f_\omega(x) := \frac{1}{J}\sum_{j=1}^J f_{\omega,j}(x)$. Here $x \in \mathbb{R}^p$ represents the model parameter to be optimized, $f_{\omega,j}(x)$ is the cost function associated with sample j at regular worker ω, and $f_\omega(x)$ is the local cost function of regular worker ω averaging on J samples. Our goal is to solve (1) in the presence of arbitrary malicious messages sent by Byzantine workers, while guarantee communication efficiency.

We make the following assumptions in the analysis.

Assumption 1 (Strong convexity and Lipschitz continuous gradients)**.** *The cost function f is μ-strong convex and has L-Lipschitz continuous gradients.*

Assumption 2 (Bounded outer variation)**.** *For any $x \in \mathbb{R}^p$, the variation of the local gradients at the regular workers w.r.t. the global gradient is upper-bounded by*

$$\frac{1}{R}\sum_{\omega\in\mathcal{R}} \|\nabla f_\omega(x) - \nabla f(x)\|^2 \le \sigma^2. \tag{2}$$

Assumption 3 (Bounded inner variation)**.** *For every regular worker $\omega \in \mathcal{R}$ and any $x \in \mathbb{R}^p$, the variation of its stochastic gradient w.r.t. its local gradient is upper-bounded by*

$$E_{i_\omega^t}\left\|\nabla f_{\omega,i_\omega^t}(x) - \nabla f_\omega(x)\right\|^2 \le \zeta^2,\ \forall\omega\in\mathcal{R}. \tag{3}$$

Assumption 4 (Bounded stochastic gradients)**.** *For every regular worker $\omega \in \mathcal{R}$ and any $x \in \mathbb{R}^p$, its stochastic gradient is upper-bounded by*

$$E_{i_\omega^t}\left\|\nabla f_{\omega,i_\omega^t}(x)\right\|^2 \le G^2,\ \forall\omega\in\mathcal{R}. \tag{4}$$

Assumption 1 is standard. Assumptions 2 and 3 bound the outer variation that describes the sample heterogeneity among the regular workers, and the inner variation that describes the sample heterogeneity on every regular worker [27]. Assumption 4 is often used to bound the compression noise [28]. Note that our proposed method does not need this assumption.

To reduce the communication cost of the federated learning system, one can compress the messages sent by the workers to the master node. Commonly used compressors are either biased or unbiased. In this paper we focus on unbiased compressors [8, 9, 22]. General, possibly biased compressors are discussed in the extended version of this paper [26].

Definition 1 (Unbiased compressor)**.** *A randomized operator $\mathcal{Q}$: $\mathbb{R}^p \to \mathbb{R}^p$ is an* unbiased compressor *if it satisfies*

$$\begin{aligned} E_{\mathcal{Q}}[\mathcal{Q}(x)] =& x, \\ E_{\mathcal{Q}}\|\mathcal{Q}(x) - x\|^2 \le& \delta\|x\|^2, \quad \forall x\in\mathbb{R}^p, \end{aligned} \tag{5}$$

where δ is a non-negative constant.

Typical unbiased compressors include randomized quantization [8] and rand-k sparsification [9]. Loosely speaking, δ can be viewed as the compression ratio. When δ approaches zero, there is little compression.

3. COMPRESSION AND STOCHASTIC NOISE IN BYZANTINE-ROBUST COMPRESSED SGD

For Byzantine-robust and communication-efficient federated learning, we first consider a vanilla approach that combines the distributed SGD with geometric median aggregation and stochastic gradient compression. We then theoretically point out that compression and stochastic noise significantly weakens its ability to tolerate Byzantine attacks.

The Byzantine-robust compressed SGD is similar to the attacks-free compressed SGD, except that the aggregation rule at the master node is changed from mean to geometric median. At iteration t, the master node broadcasts the model parameter x^t to all the workers. Then each regular worker $\omega \in \mathcal{R}$ randomly selects a sample with index i_ω^t to compute a local stochastic gradient $\nabla f_{\omega, i_\omega^t}(x^t)$. Each Byzantine worker $\omega \in \mathcal{B}$ generates an arbitrary malicious $p \times 1$ vector $*$ instead. Denote v_ω^t the vector held by each worker $\omega \in \mathcal{W}$, as

$$v_\omega^t = \begin{cases} \nabla f_{\omega, i_\omega^t}(x^t), & \omega \in \mathcal{R}, \\ *, & \omega \in \mathcal{B}. \end{cases} \tag{6}$$

To reduce the communication cost, now each worker $\omega \in \mathcal{W}$ sends the compressed message $\mathcal{Q}(v_\omega^t)$ to the master node. Upon receiving the compressed messages, the master node updates the model parameter as

$$x^{t+1} = x^t - \gamma \cdot \underset{\omega \in \mathcal{W}}{\text{geomed}} \{\mathcal{Q}(v_\omega^t)\}, \tag{7}$$

where $\gamma > 0$ is the step size and the geometric median outputs

$$\underset{\omega \in \mathcal{W}}{\text{geomed}} \{v_\omega^t\} := \arg\min_v \sum_{\omega \in \mathcal{W}} \left\| v - v_\omega^t \right\|. \tag{8}$$

Here we suppose the Byzantine workers obey the compression rule too. Otherwise, their identities are easy to recognize.

Theorem 1 (Convergence of Byzantine-robust compressed SGD)**.** *Consider the Byzantine-robust compressed SGD update* (7) *with geometric median aggregation and using an unbiased compressor. Under Assumptions 1, 2, 3, and 4, if the number of Byzantine workers satisfies $B < \frac{W}{2}$ and the step size γ satisfies $\gamma \leq \frac{\mu}{2L^2}$, then*

$$E \left\| x^t - x^* \right\|^2 \leq (1 - \gamma\mu)^t \Delta_1 + \Delta_2, \tag{9}$$

where Δ_1 is a constant and

$$\Delta_2 := \frac{4}{\mu^2} \left(C_\alpha^2 \sigma^2 + C_\alpha^2 \zeta^2 + C_\alpha^2 \delta G^2 \right), \tag{10}$$

with $\alpha := \frac{B}{W}$ and $C_\alpha := \frac{2-2\alpha}{1-2\alpha}$.

Theorem 1 asserts that the Byzantine-robust compressed SGD converges to a neighborhood of the optimal solution. The asymptotic learning error Δ_2 is linear with C_α^2, which is determined by the number of Byzantine workers. The term of δG^2 does not appear in analyzing the attacks-free compressed SGD [26] but rises here, showing the impact of compression noise in the presence of Byzantine attacks. Technically speaking, this is due to the introduction of the biased robust aggregation to defend against Byzantine attacks. Unlike the unbiased and non-robust mean aggregation, robust aggregation rules are often biased so as to handle the outliers caused by Byzantine attacks. This feature in turn amplifies the impact of compression noise on the asymptotic learning error. Thus, it is necessary to reduce the compression noise for Byzantine-robust compressed federated learning. The stochastic noise, characterized by the inner variation ζ^2 and the outer variation σ^2, also affects the asymptotic learning error. Motivated by this, we propose to reduce both compression and stochastic noise to reach a better neighborhood of the optimal solution.

4. BROADCAST: REDUCING BOTH COMPRESSION & STOCHASTIC NOISE

We start from reducing the influence of stochastic noise. Variance reduction techniques have been widely used to accelerate convergence of stochastic algorithms [23]. Motivated by the theoretical findings in Theorem 1, we combine the distributed SAGA, a popular variance reduction approach, to enhance the Byzantine-robustness. We stress that other variance reduction techniques, such as SVRG [19] and momentum [20], could be also applicable.

In the distributed SAGA, each worker stores the most recent stochastic gradient for all of its local data samples. When worker ω randomly selects a sample with index i_ω^t at iteration t, the corrected stochastic gradient is

$$\nabla f_{\omega, i_\omega^t}(x^t) - \nabla f_{\omega, i_\omega^t}(\phi_{\omega, i_\omega^t}^t) + \frac{1}{J} \sum_{j=1}^{J} \nabla f_{\omega, j}(\phi_{\omega, j}^t), \tag{11}$$

where

$$\phi_{\omega, j}^{t+1} = \begin{cases} \phi_{\omega, j}^t, & j \neq i_\omega^t, \\ x^t, & j = i_\omega^t. \end{cases} \tag{12}$$

That is to say, worker ω corrects the stochastic gradient by first subtracting the previously stored stochastic gradient of sample i_ω^t, and then adding the average of all the stored stochastic gradients of J samples. At the presence of Byzantine workers, the vector calculated at ω can be represented as

$$g_\omega^t = \begin{cases} \nabla f_{\omega, i_\omega^t}(x^t) - \nabla f_{\omega, i_\omega^t}(\phi_{\omega, i_\omega^t}^t) & \\ \qquad + \frac{1}{J} \sum_{j=1}^{J} \nabla f_{\omega, j}(\phi_{\omega, j}^t), & \omega \in \mathcal{R}, \\ \qquad *, & \omega \in \mathcal{B}, \end{cases} \tag{13}$$

where $*$ represents an arbitrary $p \times 1$ vector.

Then, we use gradient different compression to eliminate the compression noise [21, 22]. At iteration t, each worker ω and the master node maintain the same vector $h_\omega^t \in \mathbb{R}^p$, which is initialized by the same value and updated following the same rule. Each worker ω compresses the difference $g_\omega^t - h_\omega^t$, other than g_ω^t itself, and sends to the master node. After receiving the compressed difference $\mathcal{Q}(g_\omega^t - h_\omega^t)$, the master node approximates the corrected stochastic gradient as

$$\hat{g}_\omega^t = h_\omega^t + \mathcal{Q}(g_\omega^t - h_\omega^t). \tag{14}$$

Upon collecting all approximations $\hat{g}_\omega^t$, the master node updates the model parameter as

$$x^{t+1} = x^t - \gamma \cdot \underset{\omega \in \mathcal{W}}{\text{geomed}} \{\hat{g}_\omega^t\}. \tag{15}$$

With the compressed difference, each worker ω and the master node both update h_ω as

$$h_\omega^{t+1} = h_\omega^t + \beta \mathcal{Q}(g_\omega^t - h_\omega^t), \tag{16}$$

where β is a hyperparameter. Compared to directly compressing the corrected stochastic gradients, compressing the differences gradually eliminates the compression noise, as we shall see in the theoretical analysis.

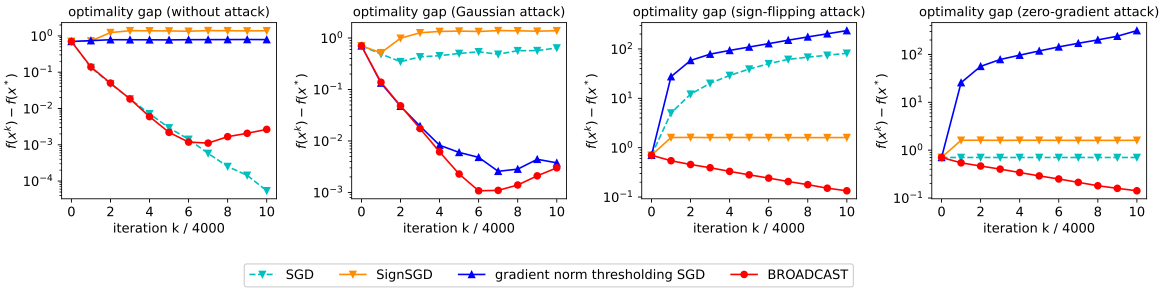

Fig. 1. Comparison between proposed algorithm and existing methods.

The proposed algorithm, Byzantine-RObust Aggregation with gradient Difference Compression And STochastic variance reduction (BROADCAST), jointly reduces the compression and stochastic noise, and is analyzed as follows.

Theorem 2 (Convergence of BROADCAST)**.** *Consider the BROADCAST update* (15) *with geometric median aggregation and using an unbiased compressor. Under Assumptions 1 and 2, if the number of Byzantine workers satisfies* $B < \frac{W}{2}$ *and* $\delta C_\alpha^2 \le \frac{\mu^2}{56L^2}$, *the hyperparameter satisfies* $\beta(1+\delta) \le 1$, *and the step size* γ *satisfies* $\gamma \le \frac{\beta\mu}{4\sqrt{35}\sqrt{1+5\delta}\cdot J^2L^2C_\alpha}$, *then*

$$E\left\|x^t - x^*\right\|^2 \le \left(1 - \frac{\gamma\mu}{2}\right)^t \Delta_1 + \Delta_2, \tag{17}$$

where Δ_1 *is a constant and*

$$\Delta_2 := \frac{140}{17\mu^2}(1+6\delta)C_\alpha^2\sigma^2. \tag{18}$$

The asymptotic learning error Δ_2 of BROADCAST is no longer dependent on the inner variation and compression noise, but from the outer variation. In contrast, the inner variation and compression noise terms both appear in the learning error of the Byzantine-robust compressed SGD. If δ is 0, meaning that no compression is applied, the learning error in BROADCAST is in the same order as that of the Byzantine-robust SAGA without compression [18]. The constant is even slightly improved by proof techniques. Thus, BROADCAST achieves gradient compression for free and achieves the same Byzantine-robustness as its uncompressed counterpart.

5. NUMERICAL EXPERIMENTS

Consider a strongly convex logistic regression problem[1]. For sample j at each regular worker ω, the sample cost is

$$f_{\omega,j}(x) = \ln(1 + \exp(-b_{\omega,j}\langle a_{\omega,j}, x\rangle)) + \frac{\xi}{2}\|x\|^2, \tag{19}$$

where $a_{\omega,j} \in \mathbb{R}^p$ is the feature, $b_{\omega,j} \in \{-1, 1\}$ is the label, and $\xi = 0.01$ is the regularization parameter. The dataset is COVTYPE with 581012 samples and $p = 54$ dimensions.

We launch $R = 50$ regular and $B = 20$ Byzantine workers. The samples are evenly and randomly allocated to the regular workers. The Byzantine attacks are Gaussian, sign-flipping and zero-gradient. For Gaussian attacks, each Byzantine worker ω obtains g_ω^t (or v_ω^t, which we will not distinguish below) from a Gaussian distribution with mean $\frac{1}{R}\sum_{\omega\in\mathcal{R}} g_\omega^t$ and variance 30. For sign-flipping attacks, each Byzantine worker ω obtains g_ω^t as $g_\omega^t = u \cdot \frac{1}{R}\sum_{\omega\in\mathcal{R}} g_\omega^t$, where the magnitude is set to $u = -3$. For zero-gradient attacks, each Byzantine worker ω obtains g_ω^t as $g_\omega^t = -\frac{1}{B}\sum_{\omega\in\mathcal{R}} g_\omega^t$ so that aggregation at the master node reaches a zero vector in the uncompressed situation. Then Byzantine worker ω compresses g_ω^t and sends to the master node. For the compressed methods, the compressor is unbiased rand-k sparsification at the regular agents, and k/p is 0.1. At the Byzantine agents we instead use biased top-k sparsification to guarantee the attacks are strong enough. The hyperparameter β in gradient difference compression is 0.1 and γ is 0.01.

Fig. 1 compares BROADCAST and existing Byzantine-robust methods with compression. SignSGD [24] transmits the signs of stochastic gradients. The gradient norm thresholding SGD [25] compresses the stochastic gradients and removes a fraction of them with the largest norms before mean aggregation. We let the fraction be 0.3, which is slightly larger than the exact fraction of Byzantine workers. With the accumulation of compression noise, SignSGD almost fails to defend all attacks and even cannot converge without attacks. For Gaussian attacks, the gradient norm thresholding SGD behaves well because all the malicious messages are removed. But it is unable to remove all the malicious messages under sign-flipping and zero-gradient attacks. In contrast, BROADCAST performs well in defending various Byzantine attacks.

6. CONCLUSIONS

Motivated by the analysis that a vanilla combination of distributed compressed SGD and geometric median aggregation suffers from compression and stochastic noise in the presence of Byzantine attacks, we develop a novel BROADCAST algorithm to reduce the noise, which consequently enhances Byzantine-robustness. Theoretical analysis and numerical experiments both validate the effectiveness of BROADCAST.

[1] Code available at https://github.com/oyhah/BROADCAST

7. REFERENCES

[1] Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

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

[3] Peter Kairouz and H Brendan McMahan, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1, pp. 1–210, 2021.

[4] Lu Zhou, Kuo-Hui Yeh, Gerhard Hancke, Zhe Liu, and Chunhua Su, "Security and privacy for the industrial internet of things: An overview of approaches to safeguarding endpoints," *IEEE Signal Processing Magazine*, vol. 35, no. 5, pp. 76–87, 2018.

[5] Sebastian U Stich, "Local SGD converges fast and communicates little," in *International Conference on Learning Representations*, 2019.

[6] Tao Lin, Sebastian U Stich, Kumar Kshitij Patel, and Martin Jaggi, "Don't use large mini-batches, use local SGD," in *International Conference on Learning Representations*, 2019.

[7] Tianyi Chen, Georgios B Giannakis, Tao Sun, and Wotao Yin, "LAG: Lazily aggregated gradient for communication-efficient distributed learning," in *Advances in Neural Information Processing Systems*, 2018, pp. 5050–5060.

[8] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Advances in Neural Information Processing Systems*, 2017, pp. 1709–1720.

[9] Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang, "Gradient sparsification for communication-efficient distributed optimization," in *Advances in Neural Information Processing Systems*, 2018, pp. 1299–1309.

[10] El Mahdi El Mhamdi, Rachid Guerraoui, and Sébastien Rouault, "The hidden vulnerability of distributed learning in Byzantium," in *International Conference on Machine Learning*, 2018, pp. 3521–3530.

[11] Yuan Chen, Soummya Kar, and Jose MF Moura, "The internet of things: Secure distributed inference," *IEEE Signal Processing Magazine*, vol. 35, no. 5, pp. 64–75, 2018.

[12] Xinyang Cao and Lifeng 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.

[13] Xinyang Cao and Lifeng Lai, "Distributed approximate Newton's method robust to byzantine attackers," *IEEE Transactions on Signal Processing*, vol. 68, pp. 6011–6025, 2020.

[14] Zhixiong Yang, Arpita Gang, and Waheed U Bajwa, "Adversary-resilient distributed and decentralized statistical inference and machine learning: An overview of recent advances under the Byzantine threat model," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 146–159, 2020.

[15] Yudong Chen, Lili Su, and Jiaming 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.

[16] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *International Conference on Machine Learning*, 2018, pp. 5650–5659.

[17] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Advances in Neural Information Processing Systems*, 2017, pp. 119–129.

[18] Zhaoxian Wu, Qing Ling, Tianyi Chen, and Georgios B Giannakis, "Federated variance-reduced stochastic gradient descent with robustness to Byzantine attacks," *IEEE Transactions on Signal Processing*, vol. 68, pp. 4583–4596, 2020.

[19] Prashant Khanduri, Saikiran Bulusu, Pranay Sharma, and Pramod K Varshney, "Byzantine resilient non-convex SVRG with distributed batch gradient computations," *arXiv preprint arXiv:1912.04531*, 2019.

[20] Sai Praneeth Karimireddy, Lie He, and Martin Jaggi, "Learning from history for Byzantine robust optimization," *arXiv preprint arXiv:2012.10333*, 2020.

[21] Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik, "Distributed learning with compressed gradient differences," *arXiv preprint arXiv:1901.09269*, 2019.

[22] Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik, "Stochastic distributed learning with gradient quantization and variance reduction," *arXiv preprint arXiv:1904.05115*, 2019.

[23] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien, "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives," in *Advances in Neural Information Processing Systems*, 2014, pp. 1646–1654.

[24] Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar, "SignSGD with majority vote is communication efficient and fault tolerant," *arXiv preprint arXiv:1810.05291*, 2018.

[25] Avishek Ghosh, Raj Kumar Maity, Swanand Kadhe, Arya Mazumdar, and Kannan Ramchandran, "Communication-efficient and Byzantine-robust distributed learning with error feedback," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 3, pp. 942–953, 2021.

[26] Heng Zhu and Qing Ling, "Broadcast: Reducing both stochastic and compression noise to robustify communication-efficient federated learning," *arXiv preprint arXiv:2104.06685*, 2021.

[27] Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu, "D2: Decentralized training over decentralized data," in *International Conference on Machine Learning*, 2018, pp. 4848–4856.

[28] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi, "Sparsified SGD with memory," in *Advances in Neural Information Processing Systems*, 2018, pp. 4447–4458.