Bridging Differential Privacy and Byzantine-Robustness via Model Aggregation

Heng Zhu[1,2] , **Qing Ling**[1]
[1]Sun Yat-sen University
[2]University of California, San Diego
hez007@ucsd.edu, lingqing556@mail.sysu.edu.cn

Abstract

This paper aims at jointly addressing two seemly conflicting issues in federated learning: differential privacy (DP) and Byzantine-robustness, which are particularly challenging when the distributed data are non-i.i.d. (independent and identically distributed). The standard DP mechanisms add noise to the transmitted messages, and entangle with robust stochastic gradient aggregation to defend against Byzantine attacks. In this paper, we decouple the two issues via robust stochastic model aggregation, in the sense that our proposed DP mechanisms and the defense against Byzantine attacks have separated influence on the learning performance. Leveraging robust stochastic model aggregation, at each iteration, each worker calculates the difference between the local model and the global one, followed by sending the element-wise signs to the master node, which enables robustness to Byzantine attacks. Further, we design two DP mechanisms to perturb the uploaded signs for the purpose of privacy preservation, and prove that they are $(\epsilon, 0)$-DP by exploiting the properties of noise distributions. With the tools of Moreau envelop and proximal point projection, we establish the convergence of the proposed algorithm when the cost function is nonconvex. We analyze the trade-off between privacy preservation and learning performance, and show that the influence of our proposed DP mechanisms is decoupled with that of robust stochastic model aggregation. Numerical experiments demonstrate the effectiveness of the proposed algorithm.

1 Introduction

Federated learning has attracted much attention from both industry and academia at the time of rapid development of distributed intelligent devices. At each iteration of a federated learning algorithm, a master node aggregates the local information sent from workers (namely, distributed intelligent devices) to update a global model. The local data of each worker are kept private, and do not need to be shared with other workers or the master node [Konečnỳ *et al.*, 2016; Kairouz and McMahan, 2021]. In the process of transmitting the local information, such as stochastic gradients or model parameters, there are several challenges to be addressed, including data privacy, robustness to malicious attacks, and communication efficiency. In this paper, we are particularly interested in the privacy and robustness issues. In a nutshell, privacy and robustness are to handle two different threats. Privacy concerns the threat from the *curious but honest* master node, who potentially expects to recover the private local data from the transmissions of local information. Robustness concerns the threat from the *dishonest and adversarial* workers, who aim at biasing the learning process.

In privacy-preserving data analysis, differential privacy (DP) [Dwork and Roth, 2014] is a gold standard and has a wide range of applications. In the popular parameter server architecture and distributed stochastic gradient descent (SGD) algorithm, adding Gaussian noise to the transmitted stochastic gradients is a common approach to achieve DP [Song *et al.*, 2013; Abadi *et al.*, 2016; Wei *et al.*, 2020]. The level of privacy guarantee is tuned by the variance of added noise. For adversarial behaviors, we consider the Byzantine attacks model to characterize that some workers may send faulty messages to bias the aggregation at the master node [Lamport *et al.*, 1982; Chen *et al.*, 2017; Mhamdi *et al.*, 2018; Kairouz and McMahan, 2021]. Byzantine attacks are devastating to the distributed SGD algorithm, in which the master node uses mean aggregation for the local stochastic gradients. A common remedy is to replace mean aggregation with other robust aggregation rules, such as geometric median, median, trimmed mean, etc [Chen *et al.*, 2017; Blanchard *et al.*, 2017; Yin *et al.*, 2018; Wu *et al.*, 2020; Karimireddy *et al.*, 2020]. However, the noise of stochastic gradients from regular workers will weaken the ability of defending against Byzantine attacks since larger variance of stochastic gradients will make the elimination of malicious messages much harder [Wu *et al.*, 2020; Karimireddy *et al.*, 2020]. Thus, the added noise of DP mechanisms shall harm the performance of robust stochastic gradient aggregation, resulting the conflict between privacy preservation and defense against Byzantine attacks. For example, the added Gaussian noise can make the regular stochastic gradients undistinguishable with the malicious messages from the Byzantine workers. The work of [Guerraoui *et al.*, 2021b] formally analyzes the incompatibility of existing robust stochastic gradient ag-

gregation rules and DP mechanisms, and [Guerraoui *et al.*, 2021a] further shows the multiplicative influence of robust stochastic gradient aggregation and DP mechanisms on the learning performance.

In this paper, we tackle the problem by decoupling the two issues via model aggregation rather than the common gradient aggregation. We propose a Differentially-Private Robust Stochastic model Aggregation (DP-RSA) algorithm to jointly address the privacy and robustness issues in federated learning, where the workers have non-i.i.d. data. At each iteration, each worker calculates the difference between the local model and the global one, followed by sending the element-wise signs to the master node, which enables defense against Byzantine attacks over non-i.i.d. data. Adaptive to robust stochastic model aggregation, we design two DP mechanisms, Sign-Flipping and Sign-Gaussian, to perturb the uploaded signs for the purpose of privacy preservation. By theoretical analysis, we display the trade-off between privacy preservation and learning performance, and point out that the separated, additive influence of our proposed DP mechanisms and robust stochastic model aggregation. Beyond provable privacy preservation and Byzantine-robustness, DP-RSA also enjoys favorable communication efficiency as the workers only send signs to the master node.

Proving privacy preservation and Byzantine-robustness for DP-RSA is challenging. To show the proposed DP mechanisms satisfy $(\epsilon, 0)$-DP, we have to investigate the impact of added noise to the signs. In particular, for the Sign-Gaussian mechanism, analyzing the impact of added Gaussian noise on the signs is difficult, and we address it by exploiting the cumulative distribution function (CDF) of Gaussian distribution. To show the convergence of DP-RSA, we must handle the nonconvex and nonsmooth cost function, for which common measures of convergence are not applicable. We leverage Moreau envelop and proximal point projection [Davis and Drusvyatskiy, 2019] to establish the convergence under the assumption of weak convexity.

Our contributions are summarized as follows.

- We propose a Differentially-Private Robust Stochastic model Aggregation (DP-RSA) algorithm for federated learning over distributed non-i.i.d. data, simultaneously meeting the requirements of privacy preservation, Byzantine-robustness, and communication efficiency.
- We design two DP mechanisms, Sign-Flipping and Sign-Gaussian, to perturb the uploaded signs for the purpose of privacy preservation. We rigorously prove that both mechanisms satisfy $(\epsilon, 0)$-DP. The proofs can be extended to other DP mechanisms that involve signs.
- We prove the convergence of DP-RSA and point out the additive impact of DP mechanisms and robust stochastic model aggregation on the learning performance. For the nonconvex and nonsmooth cost function, we leverage Moreau envelop and proximal point projection to establish the convergence. The convergence analysis is novel in the context of robust nonconvex distributed learning.

2 Problem Formulation

Consider a distributed federated learning system with one master node and K workers. 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 $\mathcal{B}$, where $K = r + b$. Note that the numbers and identities of regular and Byzantine workers are unknown to the master node. 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 acceptable solution to the nonconvex distributed learning problem

$$\min_{\tilde{x}\in\mathbb{R}^d} \sum_{k\in\mathcal{R}} E[f_k(\tilde{x}, \zeta_k)] + f_0(\tilde{x}), \tag{1}$$

where $\tilde{x} \in \mathbb{R}^d$ is the model to be optimized, $f_k(\tilde{x}, \zeta_k)$ is the nonconvex local cost function at worker k with respect to a random variable ζ_k, and $f_0(\tilde{x})$ is a regularization term at the master node. In this work we consider a non-i.i.d. setting for the distributed data. That is to say, the random variables $\zeta_k \sim D_k$, where D_k represent the data distributions at workers k and they can be different to each other at different workers. Note that non-i.i.d. data distribution is common in federated learning, and brings remarkable challenges for designing Byzantine-robust algorithms.

2.1 Differential Privacy (DP)

The definition of differential privacy (DP) is as follows.

Definition 1. *A randomized algorithm $\mathcal{M}$ is (ϵ, δ)-DP if for all $\mathcal{I}_a, \mathcal{I}_b$ that are a pair of adjacent inputs $\|\mathcal{I}_a - \mathcal{I}_b\|_1 \leq 1$ and for all possible set of outputs $\mathcal{O}$, it holds*

$$\Pr[\mathcal{M}(\mathcal{I}_a) \in \mathcal{O}] \leq e^{\epsilon}\Pr[\mathcal{M}(\mathcal{I}_b) \in \mathcal{O}] + \delta. \tag{2}$$

Here (ϵ, δ) represents the privacy budget to guarantee DP. The privacy loss ϵ controls the trade-off between privacy and utility of the algorithm, and δ is the probability that ϵ-DP can fail. In a federated learning system, at each iteration the workers send their local messages to the master node once. Thus, a local randomizer should be used to perturb the local message sent from each worker. In the context of distributed learning, we analyze the per-round-of-communication privacy budget (ϵ, δ) to ensure privacy, which is also the case in the works of, for example, [Agarwal *et al.*, 2018; Jin *et al.*, 2020; Guerraoui *et al.*, 2021b]. Indeed, it is also doable to exploit advanced composition theorems [Dwork and Roth, 2014; Kairouz *et al.*, 2015] or the analytical moments accountant method [Abadi *et al.*, 2016; Mironov, 2017] to obtain the multi-round privacy loss.

3 Algorithm Development

In this section we develop an algorithm that jointly addresses the privacy and robustness issues. We adopt the idea of robust stochastic model aggregation to handle the attacks on the non-i.i.d. data distribution, and develop DP mechanisms to protect data privacy.

3.1 Byzantine-Robustness via Model Aggregation

We first introduce robust stochastic model aggregation to defend against Byzantine attacks [Li *et al.*, 2019]. Note that (1) can be rewritten as

$$\min_{x} \sum_{k\in\mathcal{R}} E[f_k(x_k,\zeta_k)] + f_0(x_0), \quad \text{s.t. } x_0 = x_k,\ \forall k\in\mathcal{R}, \tag{3}$$

where $x := [\cdots; x_k; \cdots; x_0] \in \mathbb{R}^{(r+1)d}$ consists of r local models x_k at all regular workers and x_0 at the master node. Intuitively, the local models should be the same no matter the distributed data are non-i.i.d. or not – this is different to the local stochastic gradients. To enable robust stochastic model aggregation, we use an ℓ_1-norm penalty term to relax the constraints in (3), as

$$\min_{x} \sum_{k\in\mathcal{R}} \left(E[f_k(x_k,\zeta_k)] + \lambda \left\|x_k - x_0\right\|_1\right) + f_0(x_0), \tag{4}$$

where the ℓ_1-norm penalty parameterized by $\lambda > 0$ forces the local variables x_k to be close to x_0, but allows them to be different for the sake of tolerating Byzantine attacks.

When there are no Byzantine workers, we can apply the stochastic subgradient method with step size $\alpha^t > 0$ to solve (4). With $g_k^t := \nabla f_k(x_k^t, \zeta_k^t)$, at iteration $t+1$ we have

$$x_k^{t+1} = x_k^t - \alpha^t \left(g_k^t + \lambda \text{sign}(x_k^t - x_0^t)\right), \tag{5}$$

$$x_0^{t+1} = x_0^t - \alpha^t \left(\nabla f_0(x_0) + \lambda \sum_{k\in\mathcal{R}} \text{sign}(x_0^t - x_k^t)\right), \tag{6}$$

where the element-wise function $\text{sign}(\cdot)$ returns 1 for nonnegative input and -1 for negative input. At iteration $t+1$, the master node first broadcasts the model x_0^t to all workers. The regular workers k update their local models as (5), and send back $\text{sign}(x_0^t - x_k^t)$ to the master node. Upon receiving all local messages, the master node updates x_0^{t+1} as (6).

In the presence of Byzantine workers $j \in \mathcal{B}$, they can generate arbitrary vectors $z_j^t \in \mathbb{R}^d$ and send $\text{sign}(x_0^t - z_j^t)$ to the master node. Thus, the regular workers still update the local models as (5), but the master node updates x_0^{t+1} as

$$x_0^{t+1} = x_0^t - \alpha^t \Big(\nabla f_0(x_0) + \lambda \Big(\sum_{k\in\mathcal{R}} \text{sign}(x_0^t - x_k^t) + \sum_{j\in\mathcal{B}} \text{sign}(x_0^t - z_j^t)\Big)\Big). \tag{7}$$

As shown in [Li *et al.*, 2019], with robust stochastic model aggregation, each regular or Byzantine worker has the same impact on the model update at the master node (that is, $\alpha^k\lambda$ per element), regardless of the actual vector generated. As a consequence, the negative effects brought by the malicious messages are only relative to the number of Byzantine workers, not to the values of the malicious messages.

3.2 DP Mechanisms

However, the robust stochastic model aggregation updates (5) and (7) have the risk of leaking private data, since each regular worker k still needs to send $\text{sign}(x_0^t - x_k^t)$ to the master node. To guarantee DP, we must introduce randomness in the transmitted signs. Here we propose two DP mechanisms adaptive to robust stochastic model aggregation.

Sign-Flipping Mechanism

We first propose a straightforward mechanism to introduce randomness in the transmitted signs. The i-th element of vector $\text{sign}(x_0^t - x_k^t)$ flips its sign with probability $1-\gamma$, as

$$\text{Flip}(\text{sign}(x_0^t - x_k^t)_i) = \begin{cases} -\text{sign}(x_0^t - x_k^t)_i, & \text{with probability } 1-\gamma, \\ \text{sign}(x_0^t - x_k^t)_i, & \text{with probability } \gamma. \end{cases} \tag{8}$$

After receiving the signs from the workers, the master node cannot exactly identify the actual signs, which protects data privacy to some extent.

Sign-Gaussian Mechanism

Motivated by the conventional Gaussian mechanism in DP, we add Gaussian noise to the model difference $x_0^t - x_k^t$ and then obtain the signs. At iteration $t+1$, each regular worker k sends $\text{sign}(x_0^t - x_k^t + e_k^t)$ to the master node, where $e_k^t \sim \mathcal{N}(0, \sigma^2 I_d) \in \mathbb{R}^d$ is the multivariate Gaussian noise with zero mean and σ^2 variance, and I_d is the $d \times d$ identity matrix. Thus, the signs can be randomly changed after adding the noise. We write the Sign-Gaussian mechanism as

$$\text{Gaussian}(\text{sign}(x_0^t - x_k^t)) = \text{sign}(x_0^t - x_k^t + e_k^t). \tag{9}$$

Let $u_k^t = x_0^t - x_k^t$ and $y_k^t = \text{sign}(x_0^t - x_k^t + e_k^t)$ be the output of the Sign-Gaussian mechanism, where each element of y_k^t belongs to $\{1, -1\}$. We can obtain the probability distribution of the i-th element of y_k^t as

$$\Pr((y_k^t)_i | (u_k^t)_i) = \Phi\left(\frac{(y_k^t)_i (u_k^t)_i}{\sigma}\right), \tag{10}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution, given by $\Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-s^2/2} ds$.

3.3 DP-RSA Algorithm

Based on robust stochastic model aggregation and the two DP mechanisms, we propose the DP-RSA algorithm, which achieves privacy-preserving and Byzantine-robust federated learning over non-i.i.d. data. For notational convenience, we use $F/G(\cdot)$ to denote either the Flip function in (8) or the Gaussian function in (9).

The algorithm is described as Algorithm 1. At iteration $t+1$, the master node broadcasts the model x_0^t to all workers. The regular workers $k \in \mathcal{R}$ update their local models, and send back the perturbed signs $F/G(\text{sign}(x_0^t - x_k^t))$ to the master node. The Byzantine workers $j \in \mathcal{B}$ generate arbitrary vectors z_j^t and send $F/G(\text{sign}(x_0^t - z_j^t))$ to the master node. With particular note, for the Byzantine workers $j \in \mathcal{B}$, $F/G(\text{sign}(x_0^t - z_j^t))$ and $\text{sign}(x_0^t - z_j^t)$ are essentially equivalent. Upon receiving the signs from all workers, the master node updates x_0^{t+1} as

$$x_0^{t+1} = x_0^t - \alpha^t \Big(\nabla f_0(x_0) + \lambda \Big(\sum_{k\in\mathcal{R}} F/G(\text{sign}(x_0^t - x_k^t)) + \sum_{j\in\mathcal{B}} \text{sign}(x_0^t - z_j^t)\Big)\Big). \tag{11}$$

The robust stochastic model aggregation rule is able to defend against Byzantine attacks even for non-i.i.d. data. The

Algorithm 1 DP-RSA

Input: Step size α^t, penalty parameter λ, hyperparameter γ in Sign-Flipping or σ in Sign-Gaussian
Initialize: Initialize x_0^0 for master node

for $t = 0, 1, \dots$ **do**
  Master Node:
  Broadcast x_0^t to all workers
  Receive $F/G(\text{sign}(x_0^t - x_k^t))$ from regular workers and $F/G(\text{sign}(x_0^t - z_j^t))$ from Byzantine workers
  Update x_0^{t+1} as (11)
  Worker k **or** j:
  if $k \in \mathcal{R}$ **then**
    Receive x_0^t from master node
    Send $F/G(\text{sign}(x_0^t - x_k^t))$ to master node
    Randomly select samples $\zeta_k^t \sim D_k$ and obtain stochastic gradient $g_k^t = \nabla f_k(x_k^t, \zeta_k^t)$
    Update local model x_k^{t+1}
  else if $j \in \mathcal{B}$ **then**
    Receive x_0^t from master node
    Generate arbitrary malicious vector z_j^t
    Send $F/G(\text{sign}(x_0^t - z_j^t))$ to master node
  end if
end for

DP mechanisms can ensure data privacy during the training process. In addition, the transmissions of signs are communication-efficient. Thus, our proposed DP-RSA simultaneously meets the requirements of privacy preservation, Byzantine-robustness, and communication efficiency.

Remark 1. *It is of interest to observe that the Sign-Flipping mechanism shares similarity with Sign-Flipping Attacks, while the Sign-Gaussian mechanism is close to Gaussian Attacks. These observations indicate the trade-off between privacy preservation and learning performance. In addition, since the uploaded signs of all workers have equal contributions to the model update at the master node, the impact of attacks from Byzantine workers is not coupled with that of DP mechanisms used in regular workers. We will characterize these observations in the ensuing theoretical analysis.*

4 Theoretical Analysis

In this section, we theoretically analyze the DP-RSA algorithm. We first prove the proposed Sign-Flipping and Sign-Gaussian mechanisms satisfy $(\epsilon, 0)$-DP. Then we analyze the convergence of DP-RSA for the non-convex problem.

Now we give several assumptions used in the analysis. For notational convenience, define $f_k(\tilde{x}) := E[f_k(\tilde{x}, \zeta_k)]$ as the local cost function of regular worker k.

Assumption 1 (Weakly Convexity)**.** *The local cost functions $f_k(\tilde{x})$ of regular workers k and the regularization term $f_0(\tilde{x})$ are ρ-weakly convex, which implies $f_k(\tilde{x}) + \frac{\rho}{2}\|\tilde{x}\|^2$ and $f_0(\tilde{x}) + \frac{\rho}{2}\|\tilde{x}\|^2$ are convex. For any $\tilde{x}, \tilde{y} \in \mathbb{R}^d$, it holds*

$$f_k(\tilde{y}) \geq f_k(\tilde{x}) + \langle \nabla f_k(\tilde{x}), \tilde{y} - \tilde{x} \rangle - \frac{\rho}{2}\|\tilde{y} - \tilde{x}\|^2, \quad \forall k \in \mathcal{R} \cup 0. \tag{12}$$

Assumption 2 (Bounded Gradient)**.** *For any regular worker $k \in \mathcal{R}$ and any $x_k^t \in \mathbb{R}^d$, the stochastic gradient $g_k^t := \nabla f_k(x_k^t, \zeta_k^t)$ is upper-bounded by*

$$E\|g_k^t\|^2 \leq M^2. \tag{13}$$

For any $x_0^t \in \mathbb{R}^d$, the gradient $\nabla f_0(x_0^t)$ at the master node is also upper-bounded by

$$\|\nabla f_0(x_0^t)\|^2 \leq M^2. \tag{14}$$

Assumption 1 relaxes the assumption of convexity. In the fields of statistical learning and signal processing, weakly convex functions are common, such as nonlinear least squares, phase retrieval, robust principal component analysis, and so on. For example, a function in the form of $f_1(\cdot) + f_2(\cdot)$, where $f_1(\cdot)$ has ρ-Lipschitz continuous gradient and $f_2(\cdot)$ is closed and convex, is ρ-weakly convex [Davis and Grimmer, 2019]. Assumption 2 is common in differentially private machine learning and introduced to control the sensitivity. This assumption is natural in deep learning since gradient clipping is a standard operation to constrain the gradient norms.

4.1 DP Guarantee

For the Sign-Flipping mechanism, it is straightforward to show that it satisfies $(\epsilon, 0)$-DP; see the extended version of this paper [Zhu and Ling, 2022].

Theorem 1. *The Sign-Flipping operation given by (8) satisfies $(\ln \frac{\gamma}{1-\gamma}, 0)$-DP.*

Next, we prove that the proposed Sign-Gaussian mechanism satisfies $(\epsilon, 0)$-DP. We use u_k^t and $u_k^{t\prime}$ to represent the outputs of two adjacent datasets and $v_k^t = u_k^{t\prime} - u_k^t$. The vector v_k^t satisfies $\|v_k^t\| \leq \Delta u$, where Δu is the ℓ_2-norm sensitivity of u_k^t. To measure the privacy loss (PL) in DP, we consider

$$\begin{aligned} PL &= \ln \frac{\Pr(y_k^t | u_k^t)}{\Pr(y_k^t | u_k^t + v_k^t)} \\ &= \sum_{i=1}^{d} \ln \frac{\Phi((y_k^t)_i (u_k^t)_i / \sigma)}{\Phi((y_k^t)_i ((u_k^t)_i + (v_k^t)_i)/\sigma)} \\ &= \sum_{i=1}^{d} \left[\ln \Phi(\frac{(y_k^t)_i (u_k^t)_i}{\sigma}) - \ln \Phi(\frac{(y_k^t)_i ((u_k^t)_i + (v_k^t)_i)}{\sigma}) \right]. \end{aligned} \tag{15}$$

Observe that the function $\ln \Phi(\cdot)$ is crucial in privacy analysis. Actually its derivative is related to Mill's ratio [Sampford, 1953; Pinelis, 2019], which can be used to characterize $\ln \Phi(\cdot)$. Based on the property of Mill's ratio, we give a novel proof of the Sign-Gaussian mechanism. The proof is left to [Zhu and Ling, 2022].

Theorem 2. *If the variance σ^2 of added Gaussian noise satisfies $\sigma > \max\{\max_i \frac{2(u_k^t)_i}{3}, \frac{4\Delta u}{\epsilon}\}$, then the Sign-Gaussian operation given by* (9) *satisfies $(\epsilon, 0)$-DP, where $\epsilon \in (0, 8)$ is a constant.*

As mentioned in Definition 1, the constant δ in (ϵ, δ)-DP represents the probability that ϵ-DP fails, which is zero in our analysis. When we use the constant step size $\alpha^t = \alpha$, Δu can

be set as $2\alpha M$. Observe that as the algorithm evolves, the distance between the local and global models can be closer, resulting smaller variance of added Gaussian noise.

Note that [Jin *et al.*, 2020] proposes a similar DP-Sign operation in the SignSGD algorithm, and proves that it satisfies (ϵ, δ)-DP when the noise variance σ^2 is properly chosen. However, the proof in [Jin *et al.*, 2020] just follows that for the conventional Gaussian mechanism, and hence ensures the same level privacy guarantee. In contrast, our proof exploits the CDF of Gaussian distributions, and leads to $\delta = 0$. Our proof techniques can also be applied in analyzing other DP algorithms involving signs.

4.2 Convergence Analysis

Here we establish the convergence of DP-RSA in the nonconvex setting. Observe that now the cost function (4) is nonconvex and nonsmooth. One main challenge in the analysis is that we cannot use the optimality gap of function value or iterate for convex functions, nor the stationary condition for nonconvex smooth functions, as the measures of convergence. We use Moreau envelop and proximal point projection [Davis and Drusvyatskiy, 2019] to handle this challenge. The proof can be found in [Zhu and Ling, 2022].

For a continuous weakly convex function $h(\tilde{x})$ with $\tilde{x} \in \mathbb{R}^d$, its Moreau envelope $h_\beta(\tilde{x})$ and proximal point projection $\text{prox}_{\beta h}(\tilde{x})$ are respectively defined as

$$h_\beta(\tilde{x}) := \min_{\tilde{y}\in\mathbb{R}^d} h(\tilde{y}) + \frac{1}{2\beta}\left\|\tilde{y}-\tilde{x}\right\|^2, \tag{16}$$

$$\text{prox}_{\beta h}(\tilde{x}) := \arg\min_{\tilde{y}\in\mathbb{R}^d} h(\tilde{y}) + \frac{1}{2\beta}\left\|\tilde{y}-\tilde{x}\right\|^2. \tag{17}$$

Based on the definitions, we immediately have

$$\nabla h_\beta(\tilde{x}) = \frac{1}{\beta}(\tilde{x} - \text{prox}_{\beta h}(\tilde{x})). \tag{18}$$

More importantly, for any point $\tilde{x} \in \mathbb{R}^d$, its proximal point $\hat{x} = \text{prox}_{\beta h}(\tilde{x})$ satisfies

$$\left\|\hat{x}-\tilde{x}\right\| = \beta\left\|\nabla h_\beta(\tilde{x})\right\|, \tag{19}$$

$$\left\|\partial h(\hat{x})\right\| \le \left\|\nabla h_\beta(\tilde{x})\right\|, \tag{20}$$

where $\partial h(\hat{x})$ is any subgradient of $h(\cdot)$ at $\hat{x}$. That is to say, for a function $h(\cdot)$, a small gradient norm $\|\nabla h_\beta(\tilde{x})\|$ implies two facts: $\tilde{x}$ is close to its proximal point $\hat{x}$ and $\hat{x}$ is nearly a stationary point of $h(\cdot)$. Therefore, we can use the measure $\|\nabla h_\beta(\tilde{x})\|^2$ to establish the convergence of DP-RSA.

Before proving the convergence of DP-RSA, we further investigate the proposed DP mechanisms. With the Sign-Flipping mechanism, we have

$$\begin{aligned}\mathbb{E}_F[\text{Flip}(\text{sign}(x_0^t - x_k^t)_i)] &= \gamma\text{sign}(x_0^t - x_k^t)_i \\ &+ (1-\gamma)\text{sign}(x_k^t - x_0^t)_i,\end{aligned} \tag{21}$$

where $\mathbb{E}_F$ represents the expectation only with respect to the Sign-Flipping operation. With the Sign-Gaussian mechanism, we have

$$\begin{aligned}\mathbb{E}_G[\text{Gaussian}(\text{sign}(x_0^t - x_k^t)_i)] &= \left(\Phi(\frac{|(u_k^t)_i|}{\sigma})\right)\text{sign}(x_0^t - x_k^t)_i \\ &+ \left(1-\Phi(\frac{|(u_k^t)_i|}{\sigma})\right)\text{sign}(x_k^t - x_0^t)_i,\end{aligned} \tag{22}$$

where $\mathbb{E}_G$ represents the expectation only with respect to the Sign-Gaussian operation. To unify the convergence analysis for the two mechanisms, below we let $\tilde{\gamma} = \gamma$ in the Sign-Flipping mechanism, as well as $\gamma = \max_i \Phi(|(u_k^t)_i|/\sigma)$ and $\tilde{\gamma} = \min_i \Phi(|(u_k^t)_i|/\sigma)$ in the Sign-Gaussian mechanism. Thus, we can describe the two mechanisms as

$$\begin{aligned}\mathbb{E}[F/G(\text{sign}(x_0^t - x_k^t)_i)] &\le \gamma\text{sign}(x_0^t - x_k^t)_i \\ &+ (1-\tilde{\gamma})\text{sign}(x_k^t - x_0^t)_i.\end{aligned} \tag{23}$$

For (4), define

$$h(x) = \sum_{k\in\mathcal{R}} f_k(x_k) + f_0(x_0) + \sum_{k\in\mathcal{R}} \gamma\lambda\left\|x_k - x_0\right\|_1. \tag{24}$$

By the ρ-weak convexity of $f_k(x_k)$ and $f_0(x_0)$, $h(x)$ is also ρ-weakly convex. Further, we define $h_{1/\bar{\rho}}(x)$ as the Moreau envelop of $h(x)$ where $\bar{\rho} > 0$ is a constant. The following theorem shows that the DP-RSA iterate converges to a neighborhood of a stationary point of $h_{1/\bar{\rho}}(\cdot)$.

Theorem 3 (Convergence of DP-RSA)**.** *Suppose that Assumptions 1 and 2 hold. Set the step size of DP-RSA to* $\alpha^t = \alpha = \frac{1}{\sqrt{T}}$*. For any constant* $\bar{\rho} > \rho$*, it holds*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla h_{1/\bar{\rho}}(x^t)\right\|^2 \le \frac{\Delta_1}{\sqrt{T}} + \Delta_2, \tag{25}$$

where Δ_1, Δ_2 *are certain constants and* $\Delta_2 = O(\lambda^2[b^2 + (1-\tilde{\gamma})^2(r^2+r)])$.

According to Theorems 3, The learning errors Δ_2 is only linear to b^2, the squared number of Byzantine workers, not to the level of Byzantine attacks. If $b = 0$, i.e., there are no Byzantine attacks, the learning error Δ_2 is $O(\lambda^2(1-\tilde{\gamma})(r^2+r))$, which is due to the DP mechanisms. Because the regular workers send back the perturbed signs $F/G(\text{sign}(x_0^t - x_k^t))$, the learning error is relative to the number of regular workers r. In the Sign-Flipping mechanism, the larger the probability $1-\tilde{\gamma}$ of flipping the signs, the larger the learning error. In the Sign-Gaussian mechanism, the larger the variance σ^2 of added noise, the smaller $\tilde{\gamma}$, and then the larger the learning error. These theoretical observations align with our intuition. Therefore there is a trade-off between privacy preservation and learning performance. If we would like to have a stronger privacy guarantee in the training process, we end up with a larger learning error.

We also observe that the DP mechanisms and the robust stochastic model aggregation rule show an additive influence on the learning performance. Actually the Byzantine attacks and the DP mechanisms have similar impact, such that the DP mechanisms can be regarded as "good-will attacks". Our proposed algorithm can defend against Byzantine attacks, and also has the ability to accommodate the DP mechanisms.

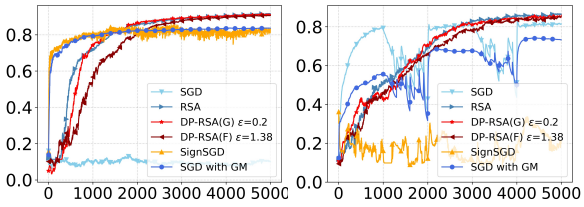
5 Numerical Experiments

We provide numerical experiments to verify the effectiveness of DP-RSA on MNIST and CIFAR10 datasets, respectively[1].

[1]The code is available at https://github.com/oyhah/DP-RSA

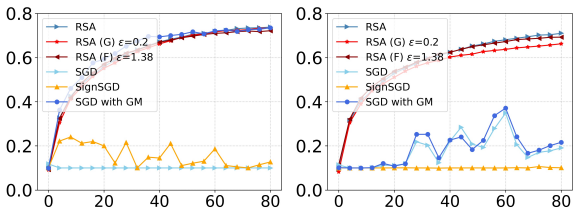
For MNIST, we train a two-layer neural network, each layer containing 50 neurons with tanh activation. In the i.i.d. setting, the 60000 training samples are evenly distributed to 30 workers. In the non-i.i.d. setting, for each digit, half of its samples are evenly distributed to 30 workers, and every 3 workers evenly share the rest half. The regularization term is $f_0(\tilde{x}) = 0.002\,\|\tilde{x}\|^2$. The penalty parameter λ is set to 0.01 and the step size α^t is set to be constant as $\alpha = 0.01$. For CIFAR10, we train a convolutional neural network (CNN) model with three fully connected layers and two convolutional layers, having about 368,000 parameters in total. In the i.i.d. setting, the 50000 training samples are evenly distributed to 20 workers. In the non-i.i.d. setting, for each class of images, half of its samples are evenly distributed to 20 workers, and every 2 workers evenly share the rest half. The regularization term is $f_0(\tilde{x}) = 0.002\,\|\tilde{x}\|^2$. The penalty parameter λ is set to 0.002 and the step size is $\alpha = 0.01$.

Denote DP-RSA with the Sign-Flipping mechanism and that with the Sign-Gaussian mechanism as DP-RSA(F) and DP-RSA(G), respectively. The privacy loss ϵ is set to 0.2, 0.4 and 1.38. We consider four baselines: SGD, SignSGD, SGD with Geometric Median (GM), and RSA. All the step sizes are the same as $\alpha = 0.01$. We consider three typical attacks. The first two are applied to the i.i.d. setting and the last one is applied to the non-i.i.d. setting. (i) Gaussian Attacks: Each Byzantine worker generates a vector where each element is from a Gaussian distribution $\mathcal{N}(0, \sigma_b^2)$, where $\sigma_b = 10000$. (ii) Sign-Flipping Attacks: Each Byzantine worker calculates the true model and multiplies it with a negative constant -5 (see [Zhu and Ling, 2022]). (iii) Sample-Duplicating Attacks: All Byzantine workers pick one regular worker, and duplicate its message.



(a) Gaussian Attacks (b) Sample-Duplicating Attacks

Figure 1: Performance comparisons on MNIST. Horizontal Axis: Number of Iterations. Vertical Axis: Accuracy.



(a) Gaussian Attacks (b) Sample-Duplicating Attacks

Figure 2: Performance comparisons on CIFAR10. Horizontal Axis: Number of Epochs. Vertical Axis: Accuracy.

Fig. 1 displays the performance of compared methods on MNIST. Here the number of Byzantine workers is $b = 3$. For both attacks, SGD has no defense abilities. DP-RSA performs similarly to RSA, suggesting that the DP mechanisms do not harm the learning performance too much and has no impact on the defense against the Byzantine attacks. Under both attacks, DP-RSA outperforms SignSGD and SGD with GM. For Sample-Duplicating Attacks in the non-i.i.d. setting, SignSGD and SGD with GM is severely influenced by the non-i.i.d. data distribution and almost fail. DP-RSA still stably finds an acceptable solution and is not influenced by the data distribution too much. For the two DP mechanisms used in DP-RSA, Sign-Flipping with large ϵ has similar performance as Sign-Gaussian with small ϵ, indicating that Sign-Gaussian is more effective in practical applications.

Fig. 2 shows the performance on CIFAR10. Here the number of Byzantine workers is $b = 2$. In the i.i.d. setting, DP-RSA also performs well and is similar to RSA, while SignSGD fails. SGD with GM can defend against Gaussian attack in i.i.d. setting. But in the non-i.i.d. setting, both SignSGD and SGD with GM fail. DP-RSA can successfully defend against the sample-duplicating attack in this case.

In the MNIST experiments, we run 2.5 epochs, and thus the samples are used for nearly 2.5 times. The overall privacy is not too different with the per-epoch privacy. In the CIFAR10 experiments, we run 80 epochs. The analytical moments accountant method can be applied to calculate the overall privacy. Due to the page limit, we leave more numerical experiments to [Zhu and Ling, 2022]. Therein, we show the performance under more attacks, and how the privacy loss ϵ, the number of Byzantine workers b, and the penalty parameter λ influence the performance of DP-RSA.

6 Conclusions

We develop a Byzantine-robust and privacy-preserving federated learning algorithm, DP-RSA, over distributed non-i.i.d. data. The messages transmitted from the workers to the master node are signs of model differences, yielding a communication-efficient implementation. We design two DP mechanisms that provably ensure data privacy. We establish the convergence of DP-RSA for the nononvex cost function and analyze the impact of Byzantine attacks and DP mechanisms on the learning performance. The numerical experiments demonstrate that the proposed DP-RSA can successfully defend against several common Byzantine attacks, for both i.i.d. and non-i.i.d. cases, and protect data privacy without sacrificing the learning performance too much.

Acknowledgements

The work of Qing Ling (corresponding author) is supported by National Natural Science Foundation of China under grant 61973324, Guangdong Basic and Applied Basic Research Foundation under grant 2021B1515020094, and Guangdong Provincial Key Laboratory of Computational Science at Sun Yat-Sen University under grant 2020B1212060032.

References

[Abadi *et al.*, 2016] Martin Abadi, Andy Chu, Ian Goodfellow, Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

[Agarwal *et al.*, 2018] Naman Agarwal, Ananda Theertha Suresh, Felix Yu, Sanjiv Kumar, and Brendan Mcmahan. cpSGD: Communication-efficient and differentially-private distributed SGD. *arXiv preprint arXiv:1805.10559*, 2018.

[Blanchard *et al.*, 2017] 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*, pages 119–129, 2017.

[Chen *et al.*, 2017] 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*, 1(2):1–25, 2017.

[Davis and Drusvyatskiy, 2019] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[Davis and Grimmer, 2019] Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.

[Dwork and Roth, 2014] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3):211–407, 2014.

[Guerraoui *et al.*, 2021a] Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, Sebastien Rouault, and John Stephan. Combining differential privacy and byzantine resilience in distributed SGD. *arXiv preprint arXiv:2110.03991*, 2021.

[Guerraoui *et al.*, 2021b] Rachid Guerraoui, Nirupam Gupta, Rafaël Pinot, Sébastien Rouault, and John Stephan. Differential privacy and byzantine resilience in sgd: Do they add up? *arXiv preprint arXiv:2102.08166*, 2021.

[Jin *et al.*, 2020] Richeng Jin, Yufan Huang, Xiaofan He, Tianfu Wu, and Huaiyu Dai. Stochastic-sign sgd for federated learning with theoretical guarantees. *arXiv preprint arXiv:2002.10940*, 2020.

[Kairouz and McMahan, 2021] Peter Kairouz and Brendan McMahan. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1):1–210, 2021.

[Kairouz *et al.*, 2015] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International Conference on Machine Learning*, pages 1376–1385, 2015.

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

[Konečnỳ *et al.*, 2016] Jakub Konečnỳ, 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.

[Lamport *et al.*, 1982] Leslie Lamport, Robert E. Shostak, and Marshall C. Pease. The Byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(3):382–401, 1982.

[Li *et al.*, 2019] Liping Li, Wei Xu, Tianyi Chen, Georgios B Giannakis, and Qing Ling. RSA: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets. In *AAAI Conference on Artificial Intelligence*, pages 1544–1551, 2019.

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

[Mironov, 2017] Ilya Mironov. Rényi differential privacy. In *IEEE Computer Security Foundations Symposium*, pages 263–275, 2017.

[Pinelis, 2019] Iosif Pinelis. Exact bounds on the inverse Mill's ratio and its derivatives. *Complex Analysis and Operator Theory*, 13(4):1643–1651, 2019.

[Sampford, 1953] Michael R Sampford. Some inequalities on Mill's ratio and related functions. *The Annals of Mathematical Statistics*, 24(1):130–132, 1953.

[Song *et al.*, 2013] Shuang Song, Kamalika Chaudhuri, and Anand Sarwate. Stochastic gradient descent with differentially private updates. In *IEEE Global Conference on Signal and Information Processing*, pages 245–248, 2013.

[Wei *et al.*, 2020] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony Quek, and Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020.

[Wu *et al.*, 2020] 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*, 68:4583–4596, 2020.

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

[Zhu and Ling, 2022] Heng Zhu and Qing Ling. Bridging differential privacy and Byzantine-robustness via model aggregation. *arXiv preprint*, 2022.