Efficient Distributed Swarm Learning for Edge Computing

Xin Fan[1], Yue Wang[2], Yan Huo[1], and Zhi Tian[2]
[1]School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing, China
[2]Department of Electrical & Computer Engineering, George Mason University, Fairfax, VA, USA
E-mails: {yhuo,fanxin}@bjtu.edu.cn, {ywang56,ztian1}@gmu.edu

***Abstract*—Federated learning (FL) methods face major challenges including communication bottleneck, data heterogeneity and security concerns in edge IoT scenarios. In this paper, inspired by the success of biological intelligence (BI) of gregarious organisms, we propose a novel edge learning approach for swarm IoT, called communication-efficient and Byzantine-robust distributed swarm learning (CB-DSL), through a holistic integration of AI-enabled stochastic gradient descent and BI-enabled particle swarm optimization. To deal with non-independent and identically distributed (non-i.i.d.) data issues and Byzantine attacks, a very small amount of global data samples are introduced in CB-DSL and shared among IoT workers, which not only alleviates the local data heterogeneity effectively but also enables to fully utilize the exploration-exploitation mechanism of swarm intelligence. Further, we provide convergence analysis to theoretically demonstrate that the proposed CB-DSL is superior to the standard FL with better convergence behavior. In addition, to measure the effectiveness of the introduction of the globally shared dataset, we also evaluate the model divergence by deriving its upper bound. Numerical results verify that the proposed CB-DSL outperforms the existing benchmarks in terms of faster convergence speed, higher convergent accuracy, lower communication cost, and better robustness against non-i.i.d. data and Byzantine attacks[1].**

***Index Terms*—Distributed swarm learning, federated learning, particle swarm optimization, non-i.i.d. data, convergence analysis, model divergence analysis.**

I. Introduction

Federated learning (FL) has recently attracted great attention and resulted in fruitful attempts for learning-based applications among multiple distributed workers such as personal mobile phones, which allows distributed learning from local data without raw data exchange [1]–[3]. Standard FL methods are originally designed for ideal learning settings and wireless environments, which however face several challenges when being adopted for distributed learning among massive edge Internet of Things (IoT) devices that are usually equipped with limited capability and resources [4], [5]. As the number of model parameters goes very large in deep neural networks, transmission of all the local model updates in FL between IoT devices (working as local workers) and the parameter server (PS) incurs high communication overhead. Further, stochastic gradient descent (SGD) is widely applied for model training in FL [6], where independent and identically distributed (i.i.d.) data samples are assumed at local workers and transmission is assumed error-free in order to ensure unbiased estimates and good empirical performances [7], [8]. However, in edge IoT scenarios, local training data samples at different IoT workers turn to be statistically heterogeneous worker-by-worker, giving rise to the non-i.i.d. data issue that may considerably degrades the learning performance of standard FL methods, e.g., Federated Averaging (FedAvg) [9]. In addition, gradient-based algorithms are subject to local optimum traps in solving non-convex problems [10], [11]. This issue is aggravated in distributed settings, especially when local workers only collect small-volume data. Last but not the least, standard FL performs well in attack-free network settings, but is vulnerable to Byzantine attacks that may exist in practical edge networks [12]–[15].

Although some of the aforementioned challenges have been recently investigated in the literature of FL for edge networks and IoT applications [16]–[20], they mainly focus on the modification and customization of the standard FL techniques, which however largely neglect some important and unique characteristics of IoT devices in edge networks. Such unique characteristics include the large population of devices for many IoT applications, limited communication bandwidth available in edge networks, and non-i.i.d. local data with small data volume at individual IoT workers. By ignoring these characteristics, existing efforts on edge learning fail to consider these limitations in the learning algorithm design for edge IoT systems, which results in learning performance degradation of FL applied to practical IoT edge networks. On the other hand, biological organisms in nature have demonstrated swarm intelligence with superior strength in collectively processing information, making decisions, dealing with uncertainties, and recovering from errors and failures, even though they are individually weak. All these attributes of biological intelligence (BI) are desired by IoT edge learning systems. Notably, bio-inspired swarm optimization techniques are good at collaboratively finding the globally optimal solutions to complex optimization problems thanks to their built-in exploration-exploitation mechanism in swarms, but their convergence speed is typically slow [21].

Motivated to bridge these gaps, this paper leverages both AI and BI to develop a communication-efficient and Byzantine-robust distributed swarm learning (CB-DSL) approach, by reformulating the bio-inspired particle swarm optimization (PSO) problem as a distributed learning problem with non-i.i.d. local data and in the presence of malicious attacks. For non-convex problems, by taking advantage of the exploration-exploitation mechanism of PSO [22], our CB-DSL solutions have an increased chance to jump out of local optimum traps

[1]Our code can be found at:https://github.com/fuanxiyin/CB-DSL.git.

ICC 2023 - IEEE International Conference on Communications | 978-1-5386-7462-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICC45041.2023.10279508

via swarm intelligence. For the communication bottleneck challenge, our CB-DSL only requires the best worker having the minimum loss function value to upload its local model to the PS, which thus dramatically reduces the communication overhead and energy consumption in edge networks. To alleviate the non-i.i.d. data issue, we propose to introduce a small-volume global dataset that is shared among all local workers for dual purposes. A part of this globally shared dataset is used for training, whose effectiveness in relieving the non-i.i.d. problem is evaluated through the model divergence analysis. The other part of the global dataset is used to calculate the fair-value loss for scoring the local models. It helps to identify the per-worker best model for best worker selection, and enables to verify the uploaded local model by which the PS can screen Byzantine attackers. Our main contributions are summarized as follows.

- We propose a new CB-DSL framework by developing a holistic integration of AI-driven SGD and BI-driven PSO, to effectively handle the high communication costs, non-i.i.d. issues, non-convex problems and Byzantine attacks without sacrifice convergence speed, which cannot be achieved by SGD or PSO alone. CB-DSL offers a new paradigm of efficient and robust edge learning tailored for massive smart IoT devices in edge networks.
- From the theoretical point of view, we are the first one to systematically analyze the combination of FL and PSO, by deriving a closed-form expression to quantify the expected convergence rate achieved by our CB-DSL. Our analytical results not only reflect the impact of different settings and parameters of our CB-DSL on the performance of edge learning, but also indicate that our CB-DSL outperforms the standard FL methods such as FedAvg in terms of better convergence rate.
- We further investigate the non-i.i.d. data issue by providing a model divergence analysis to evaluate how a globally shared dataset improves the learning performance of our CB-DSL. Our theoretical result reveals that the model divergence is subject to an upper bound, which is decided by the earth mover's distance (EMD) between the data distribution at local workers and the population distribution for the whole datasets.
- Through comprehensive experiments, we test the proposed CB-DSL approach in solving image classification problems by using the MNIST dataset. Simulation results show that our CB-DSL outperforms the benchmark methods in terms of achieving the highest testing accuracy with the fastest convergence under non-i.i.d. cases and even in the presence of Byzantine attacks.

II. Distributed Swarm Learning

Consider a distributed learning model with one PS and U IoT workers, where U is very large but each worker has data of small volume in edge IoT scenarios. Assume that each worker has K_i data samples in its local dataset $\mathfrak{D}_i$, with $|\mathfrak{D}_i| = K_i$, and $i = 1, \ldots, U$. Denote $(\mathbf{x}_{i,k}, y_{i,k})$ as the k-th data sample of the i-th local worker. Let $f(\mathbf{w}; \mathbf{x}_{i,k}, y_{i,k})$ represent the loss function associated with each data sample $(\mathbf{x}_{i,k}, y_{i,k})$, where $\mathbf{w} = [w^1, \ldots, w^D]$ of size D consists of the parameters of a common learning model. The corresponding population loss function for the whole datasets $\mathfrak{D}$ and that for the local dataset $\mathfrak{D}_i$ of the i-th worker are denoted as $F(\mathbf{w}) := \mathbb{E}_{\mathfrak{D}}[f(\mathbf{w}; \mathbf{x}_{i,k}, y_{i,k})]$ and $F_i(\mathbf{w}) := \mathbb{E}_{\mathfrak{D}_i}[f(\mathbf{w}; \mathbf{x}_{i,k}, y_{i,k})]$, respectively, where $\mathfrak{D} = \bigcup_i \mathfrak{D}_i$. For distributed learning, local workers collaboratively learn $\mathbf{w}$ by minimizing

$$\textbf{P1:} \quad \mathbf{w}_i^* = \arg\min_{\mathbf{w}_i} F_i(\mathbf{w}_i), \qquad \text{s.t.,} \quad \mathbf{w}_i = \mathbf{z}, \quad \forall i, \tag{1}$$

where $\mathbf{z}$ is an auxiliary variable to enforce consensus through collaboration among distributed local workers.

A. Federated Learning

For standard FL designed in ideal learning settings and network environments, the minimization of $F_i(\mathbf{w})$ is typically carried out by the stochastic gradient descent (SGD) algorithm [6], where local workers iteratively update their local models in FL as

$$\mathbf{w}_{i,t+1} = \mathbf{w}_{i,t} - \tfrac{\alpha}{U} \textstyle\sum_{j=1}^{U} \nabla F_j(\mathbf{w}_t; \mathbf{x}_{j,k}, y_{j,k}), \tag{2}$$

where α is the learning rate and $\nabla F_j(\mathbf{w}_t; \mathbf{x}_{j,k}, y_{j,k}) = \mathbb{E}_{\mathfrak{D}_j}\left[\frac{\sum_{\mathfrak{B}_j} \nabla f(\mathbf{w}_t; \mathbf{x}_{j,k}, y_{j,k})}{|\mathfrak{B}_j|}\right]$ is the local gradient computed at each local worker using its randomly selected mini-batch $\mathfrak{B}_j \subset \mathfrak{D}_j$ with the mini-batch size $|\mathfrak{B}_j|$.

Note that (2) is the mathematical illustration of the iterative local model update, whereas the second term of global gradient averaging therein is typically implemented at the PS and then sent back to local workers. Hence, communications take place in every iteration until convergence, during which the communication overhead to acquire the sum of all U local gradients in (2) would be huge especially when U and D are large. Moreover, for complicated non-convex problems, distributed gradient-based FL solutions may converge to undesired local optima and there is unfortunately a lack of effective mechanisms to escape these traps.

B. Particle Swarm Optimization

As a bio-inspired algorithm, PSO is a stochastic optimization approach based on the movement of particles (workers) and the collaboration of swarms to iteratively and cooperatively search for an optimal solution to general optimization problems [22], [23]. The loss function in PSO is assumed to be globally common to all particles, i.e., $F_i(\cdot) = F(\cdot), \forall i$ in the problem **P1** in (1). This is however not the case in distributed leaning where $F_i(\cdot)$ is data-dependent and different worker-by-worker, which will be explained in the next subsection.

In PSO, a swarm consists of a large set of particles, $i = 1, 2..., U$. At the current iteration, the position $\mathbf{w}_{i,t}$ of each particle i presents a possible solution to the problem, and meanwhile the velocity $\mathbf{v}_{i,t}$ of each particle i denotes the updating direction for the next step. To find the globally optimal value of $F(\cdot)$, particles collaborate with each other to update their velocities and positions in an iterative manner

$$\mathbf{v}_{i,t+1} = c_0 \mathbf{v}_{i,t} + c_1(\mathbf{w}_{i,t}^p - \mathbf{w}_{i,t}) + c_2(\mathbf{w}_t^g - \mathbf{w}_{i,t}), \tag{3}$$

$$\mathbf{w}_{i,t+1} = \mathbf{w}_{i,t} + \mathbf{v}_{i,t+1}, \tag{4}$$

where the velocity is updated as a combination of three sub-directions: inertia $\mathbf{v}_{i,t}$ of the previous updating direction,

individual direction towards each particle's own historical best parameter $\mathbf{w}_{i,t}^p = \text{argmin}_{\tau=1,\cdots,t} F(\mathbf{w}_{i,\tau})$, and social direction towards the globally best parameter found by the entire swarm $\mathbf{w}_t^g = \text{argmin}_{i=1,\cdots,U} F(\mathbf{w}_{i,t}^p)$. The inertia weight c_0 is a positive number, while c_1 and c_2 are positive and random (say, uniformly distributed as $c_1 \sim \mathcal{U}(0, \delta_{c_1})$, and $c_2 \sim \mathcal{U}(0, \delta_{c_2})$) for stochastic optimization.

C. Communication-efficient and Byzantine-robust Distributed Swarm Learning

A major challenge from optimization problems to learning problems with distributed data is the lack of a common $F(\cdot)$ for global assessment, which however becomes $F_i(\cdot; \mathfrak{D}_i)$ dependent on local dataset $\mathfrak{D}_i$ in distributed learning. Facing this challenge, we first introduce a very small amount of global dataset[2]: $\mathfrak{D}^G = \mathfrak{D}_{tr}^G \cup \mathfrak{D}_{sc}^G$ to be shared by all workers, and then propose a novel edge learning framework called communication-efficient and Byzantine-robust distributed swarm learning (CB-DSL). The CB-DSL algorithm is implemented in **Algorithm 1**, and schematically illustrated through the following iterative model updating steps.

At the local workers $i = 1, \cdots, U$, the model parameters are updated in a way of integrating BI-enabled PSO with AI-enabled SGD

$$\mathbf{w}_{i,t+1} = \mathbf{w}_{i,t} + \underbrace{c_0 \mathbf{v}_{i,t} + c_1(\mathbf{w}_{i,t}^p - \mathbf{w}_{i,t}) + c_2(\mathbf{w}_t^g - \mathbf{w}_{i,t})}_{\textbf{BI}} \underbrace{-\alpha \nabla F_i(\mathbf{w}_{i,t}; \mathfrak{D}_i \cup \mathfrak{D}_{tr}^G)}_{\textbf{AI}}, \quad (5)$$

where $\mathfrak{D}_{tr}^G$ is a part of $\mathfrak{D}^G$ and used for training to relieve the non-i.i.d. problem.

Then, the local workers calculate their own historical minimum loss function values and maintain their own historical best model parameters

$$\{F_{i,t+1}^p, \mathbf{w}_{i,t+1}^p\} = \arg\min_{\tau=1,\cdots,t+1} F_i(\mathbf{w}_{i,\tau}, \mathfrak{D}_{sc}^G), \quad (6)$$

where $\mathfrak{D}_{sc}^G$ is the other part of $\mathfrak{D}^G$ and used to provide fair-value scores of local models for best-worker selection by assessing the per-worker $F_{i,t+1}^p$. Then, all workers report their $F_{i,t+1}^p$ to the PS.

Comparing the received $\{F_{i,t+1}^p\}_i$ from all local workers, the PS selects the best worker $i_{t+1}^\star$ with the global optimum function value

$$\{i_{t+1}^\star, F_{t+1}^g\} = \arg\min_{i=1,\cdots,U} F_{i,t+1}^p. \quad (7)$$

If $F_{t+1}^g < F_t^g$, then the worker with the selected index $i_{t+1}^\star$ is invited to upload its $\mathbf{w}_{i_{t+1}^\star,t+1}^p$ to the PS as the globally best model parameter $\mathbf{w}_{t+1}^g = \mathbf{w}_{i_{t+1}^\star,t+1}^p$. If $F_{t+1}^g \geq F_t^g$, then no worker is invited to upload local model parameter and the PS simply maintains the globally best model parameter and the globally best loss function value from the previous iteration as $\mathbf{w}_{t+1}^g = \mathbf{w}_t^g$ and $F_{t+1}^g = F_t^g$.

Upon receiving $\mathbf{w}_{i_{t+1}^\star,t+1}^p$ from the invited worker, the PS further uses $\mathfrak{D}_{sc}^G$ to verify the reported model parameter. If $F(\mathbf{w}_{i_{t+1}^\star,t+1}^p, \mathfrak{D}_{sc}^G) \neq F_{t+1}^g$, then a Byzantine attack is identified and the attacker is filtered out; the PS will inquire the next best local worker, until confirmed.

Communication Efficiency. Note that our CB-DSL requires U workers to share their function value $F_{i,t+1}^p$ which is only a scalar, and then invites only one local worker with the global minimum loss function value calculated using $\mathfrak{D}_{sc}^G$ to report its model parameter to the PS. Thus, our CB-DSL can dramatically reduce the overall communication overhead and energy consumption in edge networks.

Byzantine Robustness. In the process of collecting $F_{i,t+1}^p$'s from local workers, a malicious worker may send a fake $\bar{F}_{i,t+1}^p$ $(< F_{i,t+1}^p)$ to fool the PS to invite the attacker to upload its fake model parameter as the global optimum, which will undermine edge learning. Thanks to $\mathfrak{D}_{sc}^G$ in our CB-DSL, it enables the PS to screen and remove the potential Byzantine attackers, resulting our Byzantine-robust CB-DSL.

Algorithm 1 CB-DSL

Initialization:
$\mathbf{w}_{i,0}^p = \mathbf{w}_{i,0}$, $F_{i,0}^p = F_i(\mathbf{w}_{i,0}, \mathfrak{D}_{sc}^G)$, $\forall i$;
1: **for** each iteration $t = 1 : T$ **do**
2: **at the local workers:**
3: update the local model parameter $\mathbf{w}_{i,t+1}$ via (5);
4: calculate the historical minimum loss function value $F_{i,t+1}^p$ and maintain the corresponding historical best model parameter $\mathbf{w}_{i,t+1}^p$ via (6);
5: send the scalar function value $F_{i,t+1}^p$ to the PS;
6: only the invited local worker sends $\mathbf{w}_{i,t+1}^p$ to the PS;
7: **at the PS:**
8: compare the received $F_{i,t+1}^p$'s, select the best worker $i_{t+1}^\star$ and identify its function value as F_{t+1}^g via (7);
9: *if* $F_{t+1}^g < F_t^g$, *then* invite the selected worker $i_{t+1}^\star$ to upload its model parameter as the globally best model parameter $\mathbf{w}_{t+1}^g = \mathbf{w}_{i_{t+1}^\star,t+1}^p$;
10: *else*, no worker is invited and maintain the globally best model parameter and function value from the previous iteration as $\mathbf{w}_{t+1}^g = \mathbf{w}_t^g$ and $F_{t+1}^g = F_t^g$;
11: given $\mathbf{w}_{i_{t+1}^\star,t+1}^p$ received from the invited worker, verify $F(\mathbf{w}_{i_{t+1}^\star,t+1}^p, \mathfrak{D}_{sc}^G) == F_{t+1}^g$;
12: *if* an attacker is identified by $F(\mathbf{w}_{i_{t+1}^\star,t+1}^p, \mathfrak{D}_{sc}^G) \neq F_{t+1}^g$, remove it and repeat line 8 until a legitimate worker is selected.
13: **end for**

III. Convergence Analysis

In this section, we first make some definitions and assumptions for convergence analysis. With these preliminaries, the convergence behavior of our CB-DSL approach is theoretically evaluated by deriving an upper bound of the convergence rate.

[2]For the implementation point of view, a small amount (e.g., 1% of all datasets is adequate as used in our simulations) of globally shared dataset can be generated by a generative adversarial network module for keeping the privacy of workers' own local data [24], which can be either pre-stored in the IoT devices or broadcasted from the PS to all the local workers. The required resources in sharing and local storage are quite low.

A. Assumption and Definition

Assumption 1. (Lipschitz continuity, smoothness): The gradient $\nabla F_i(\mathbf{w})$ of the loss function $F_i(\mathbf{w})$ at node i is uniformly Lipschitz continuous with respect to $\mathbf{w}$, that is,

$$\|\nabla F_i(\mathbf{w}_{i,t+1}) - \nabla F_i(\mathbf{w}_{i,t})\| \leq L\|\mathbf{w}_{i,t+1} - \mathbf{w}_{i,t}\|, \ \forall i, t, \tag{8}$$

where L is a positive constant, referred as the Lipschitz constant for the loss function $F_i(\cdot)$ [3].

To facilitate analyses, we rewrite $\mathbf{w}_{i,t}^p$ and $\mathbf{w}_t^g$ in (5) as

$$\mathbf{w}_{i,t}^p = \mathbf{w}_{i,t-1} + \mathbf{v}_{i,t}^p, \tag{9}$$

$$\mathbf{w}_t^g = \mathbf{w}_{i,t-1} + \mathbf{v}_t^g, \tag{10}$$

where $\mathbf{v}_{i,t}^p$ and $\mathbf{v}_t^g$ denote the per-worker and globally optimal velocities currently used at the i-th worker.

Then, the DSL velocity update $\mathbf{v}_{i,t+1} = \mathbf{BI} + \mathbf{AI} = \mathbf{w}_{i,t+1} - \mathbf{w}_{i,t}$ in (5) can be rewritten as

$$\begin{aligned}\mathbf{v}_{i,t+1} &= c_0\mathbf{v}_{i,t} + c_1(\mathbf{v}_{i,t}^p - (\mathbf{w}_{i,t} - \mathbf{w}_{i,t-1})) \\ &\quad + c_2(\mathbf{v}_t^g - (\mathbf{w}_{i,t} - \mathbf{w}_{i,t-1})) - \alpha\nabla F_i(\mathbf{w}_{i,t}) \\ &= c_0\mathbf{v}_{i,t} + c_1(\mathbf{v}_{i,t}^p - \mathbf{v}_{i,t}) + c_2(\mathbf{v}_t^g - \mathbf{v}_{i,t}) - \alpha\nabla F_i(\mathbf{w}_{i,t}) \\ &= (c_0 - c_1 - c_2)\mathbf{v}_{i,t} + c_1\mathbf{v}_{i,t}^p + c_2\mathbf{v}_t^g - \alpha\nabla F_i(\mathbf{w}_{i,t}), \end{aligned} \tag{11}$$

where we replace $\nabla F_i(\mathbf{w}_{i,t}; \mathfrak{D}_i \cup \mathfrak{D}_{tr}^G)$ by $\nabla F_i(\mathbf{w}_{i,t})$ hereafter for symbol simplicity.

We use $\theta_{i,t}$, $\theta_{i,t}^p$, and θ_t^g to denote the angles between $\mathbf{v}_{i,t}$ and $-\nabla F_i(\mathbf{w}_{i,t})$, between $\mathbf{v}_{i,t}^p$ and $-\nabla F_i(\mathbf{w}_{i,t})$, and between $\mathbf{v}_t^g$ and $-\nabla F_i(\mathbf{w}_{i,t})$, for any i and t, respectively. Then we have $\cos\theta_{i,t} \triangleq \frac{\langle \mathbf{v}_{i,t}, -\nabla F_i(\mathbf{w}_{i,t})\rangle}{\|\mathbf{v}_{i,t}\|\|\nabla F_i(\mathbf{w}_{i,t})\|}$, $\cos\theta_{i,t}^p \triangleq \frac{\langle \mathbf{v}_{i,t}^p, -\nabla F_i(\mathbf{w}_{i,t})\rangle}{\|\mathbf{v}_{i,t}^p\|\|\nabla F_i(\mathbf{w}_{i,t})\|}$, $\cos\theta_t^g \triangleq \frac{\langle \mathbf{v}_t^g, -\nabla F_i(\mathbf{w}_{i,t})\rangle}{\|\mathbf{v}_t^g\|\|\nabla F_i(\mathbf{w}_{i,t})\|}$, $\forall i, t$.

We further assume that the above cosine-similarity measures are bounded, whose lower and upper bounds are denoted as $\underline{q} \leq \cos\theta_{i,t} \leq \overline{q}, \underline{q}^p \leq \cos\theta_{i,t}^p \leq \overline{q}^p, \underline{q}^g \leq \cos\theta_t^g \leq \overline{q}^g$, $\underline{u} \leq \frac{\|\mathbf{v}_{i,t}\|}{\|\nabla F_i(\mathbf{w}_{i,t})\|} \leq \overline{u}, \underline{u}^p \leq \frac{\|\mathbf{v}_{i,t}^p\|}{\|\nabla F_i(\mathbf{w}_{i,t})\|} \leq \overline{u}^p, \underline{u}^g \leq \frac{\|\mathbf{v}_t^g\|}{\|\nabla F_i(\mathbf{w}_{i,t})\|} \leq \overline{u}^g, \ \forall i, t$.

B. Convergence Bound

With the assumptions and definitions presented in Subsection IV.A, the convergence errors of the CB-DSL algorithm are bounded by the following **Theorem 1**.

Theorem 1. *For T communication rounds, the expected convergence rate at each worker in CB-DSL is bounded by*

$$\mathbb{E}\left[\sum_{t=1}^{T}\frac{\|\nabla F_i(\mathbf{w}_{i,t})\|^2}{T}\right] \leq \frac{F(\mathbf{w}_{i,0}) - F(\mathbf{w}^*)}{T\Phi_E}, \ \forall i, \tag{12}$$

where $\Phi_E = \alpha - \frac{2c_0 - \delta_{c_1} - \delta_{c_2}}{2}\underline{q}\underline{u} - \frac{\delta_{c_1}}{2}\overline{u}^p\overline{q}^p - \frac{\delta_{c_2}}{2}\overline{u}^g\overline{q}^g - 2L((c_0^2 - \delta_{c_1}c_0 - \delta_{c_2}c_0 + \frac{\delta_{c_1}^2}{3} + \frac{\delta_{c_2}^2}{3} + \frac{\delta_{c_1}\delta_{c_2}}{2})\overline{u}^2 + \frac{\delta_{c_1}^2}{3}(\overline{u}^p)^2 + \frac{\delta_{c_2}^2}{3}(\overline{u}^g)^2 + \alpha^2)$.

Proof: Please refer to our journal version [25]. ■

Remark 1. When c_0, δ_{c_1}, and δ_{c_2} are all set to be 0, we have $\Phi_E = \alpha - 2L\alpha^2$ in (12), and CB-DSL degenerates into FedAvg. As $\Phi_E - (\alpha - 2L\alpha^2) = \frac{\delta_{c_1} + \delta_{c_2} - 2c_0}{2}\underline{q}\underline{u} + 2L((\delta_{c_1}c_0 + \delta_{c_2}c_0 - c_0^2 - \frac{\delta_{c_1}^2}{3} - \frac{\delta_{c_2}^2}{3} - \frac{\delta_{c_1}\delta_{c_2}}{2})\overline{u}^2 - \frac{\delta_{c_1}^2}{3}(\overline{u}^p)^2 - \frac{\delta_{c_2}^2}{3}(\overline{u}^g)^2) - \frac{\delta_{c_1}}{2}\overline{u}^p\overline{q}^p - \frac{\delta_{c_2}}{2}\overline{u}^g\overline{q}^g > 0$, CB-DSL converges faster than FedAvg.

IV. Model Divergence Analysis for the Case of Non-i.i.d. Data

Consider a C-class classification problem defined over a compact space $\mathcal{X}$ and a label space $\mathcal{Y}$. The k-th data point $(\mathbf{x}_{i,k}, y_{i,k})$ on the i-th local worker distributes over $\mathcal{X} \times \mathcal{Y}$ following the distribution p_i. For the purpose of model divergence analysis, suppose a genie worker who has the population data that reflect the population distribution p of all local workers that may differ from p_i. The genie worker uses such knowledge of p to search for the globally optimal solution to the learning model, which serves as the reference to calibrate the model divergence due to the distributed non-i.i.d. data. Then the original population loss function $F(\mathbf{w}) := \mathbb{E}_{\mathfrak{D}}[f(\mathbf{w}; \mathbf{x}_{i,k}, y_{i,k})]$ can be rewritten as

$$F(\mathbf{w}) = \sum_{c=1}^{C} p(y=c)\mathbb{E}_{\mathbf{x}|y=c}[f_c(\mathbf{x}, \mathbf{w})], \tag{13}$$

where f_c denotes the probability for the c-th class, $c \in \{1, C\}$.

Then, the learning problem at the genie worker can be formulated as

$$\textbf{P2:} \quad \mathbf{w}^* = \arg\min_{\mathbf{w}} \quad \sum_{c=1}^{C} p(y=c)\mathbb{E}_{\mathbf{x}|y=c}[f_c(\mathbf{x}, \mathbf{w})]. \tag{14}$$

By solving **P2**, the model obtained at the genie worker plays as the globally optimal position in each communication round of CB-DSL. Then according to (11), the velocity at the genie worker in the $(t+1)$-th communication round is updated via

$$\mathbf{v}_{t+1}^g = c_0\mathbf{v}_t^g - \alpha\nabla F(\mathbf{w}_t^g). \tag{15}$$

The model parameter at the genie worker in the $(t+1)$-th communication round is updated as

$$\mathbf{w}_{t+1}^g = \mathbf{w}_t^g + \mathbf{v}_{t+1}^g. \tag{16}$$

Given (5) and (16), the model divergence between the i-th local worker and the genie worker is defined as

$$model \ \ divergence = \frac{\|\mathbf{w}_{i,t+1} - \mathbf{w}_{t+1}^g\|}{\|\mathbf{w}_{t+1}^g\|}. \tag{17}$$

Next, we provide **Theorem 2** to evaluate the model divergence by deriving its upper bound theoretically.

Theorem 2. *Under the assumption that $\nabla\mathbb{E}_{\mathbf{x}|y=c}[f_c(\mathbf{x}, \mathbf{w})]$ is L_c-Lipschitz for each class $c \in \{1, C\}$, we have the following inequality for the model divergence as*

$$\begin{aligned}\|\mathbf{w}_{i,t+1} - \mathbf{w}_{t+1}^g\| &\leq \beta^{t+1}\|\mathbf{w}_{i,0} - \mathbf{w}_0^g\| \\ &\quad + |c_0 - c_1 - c_2|\sum_{j=0}^{t}\beta^{t-j}\|\mathbf{v}_{i,j} - \mathbf{v}_j^g\| \\ &\quad + \alpha\sum_{c=1}^{C}\|p_i(y=c) - p(y=c)\|\sum_{j=0}^{t} f_{max}(\mathbf{w}_j^g), \end{aligned} \tag{18}$$

where $\beta = 1 + \alpha\sum_{c=1}^{C} p_i(y=c)L_c$ *and* $f_{max}(\mathbf{w}_j^g) = \max\{\nabla\mathbb{E}_{\mathbf{x}|y=c}[f_c(\mathbf{x}, \mathbf{w}_j^g)]\}_{c=1}^{C}$.

Proof: Please refer to our journal version [25]. ■

Remark 2. In (18), the initial model divergence (first term) and the velocity divergence (second term) after $(t+1)$ communication rounds are iteratively amplified by β. Since $\beta > 1$, if different local workers start from different initial model

parameters in the standard FL, then the model divergence will still be enlarged, even though the local workers have i.i.d. data.

Remark 3. In (18), the third term $\sum_{c=1}^{C} \|p_i(y=c) - p(y=c)\|$ is the EMD between the data distribution on the i-th local worker and the population distribution [26], when the distance metric is defined as $\|p_i(y=c) - p(y=c)\|$. The impact of EMD is affected by the learning rate α, the number of communication rounds t, and the class-wise maximum gradient $f_{max}(\mathbf{w}_j)$.

V. Experimental Results

This section demonstrates that our CB-DSL outperforms the benchmark methods, with better learning performance and faster convergence speed, on non-i.i.d. settings, even in the presence of Byzantine attacks.

A. System and Dataset Setting

We perform empirical simulations by conducting a handwritten-digit classification task based on the widely-used MNIST dataset[3]. We set the total number of local workers to be $U = 50$. To build the non-i.i.d. data setting upon the MNIST dataset, we first sort all the 60000 training samples based on the classification labels. Then we divide the 60000 training samples into 200 shards, each of which consists 300 samples, that are highly non-i.i.d. shard by shard [6]. We randomly allocate two shards to each local worker. The globally shared scoring dataset $\mathfrak{D}_{sc}^{G}$ consists of 2000 data samples, and the globally shared training dataset $\mathfrak{D}_{tr}^{G}$ consists of 600 data samples.

B. Different Approaches

We compare the proposed CB-DSL with FedAvg [6], given either i.i.d. or non-i.i.d. data, for different cases of globally shared dataset, including: (1) *FedAvg without any globally shared dataset* $\mathfrak{D}^{G}$. (2) *CB-DSL without any globally shared dataset* $\mathfrak{D}^{G}$: the local workers use their own local dataset to calculate $F_{i,t}^{p}$. (3) *CB-DSL with a globally shared dataset for scoring* $\mathfrak{D}_{sc}^{G}$: the local workers use the globally shared scoring dataset to calculate $F_{i,t}^{p}$ in CB-DSL. (4) *FedAvg with a globally shared dataset for training* $\mathfrak{D}_{tr}^{G}$: the local workers use both their own local dataset and the globally shared training dataset to train their local models in standard FedAvg [6]. (5) *CB-DSL with a globally shared dataset for both training* $\mathfrak{D}_{tr}^{G}$ *and scoring* $\mathfrak{D}_{sc}^{G}$: the local workers use both their own local dataset and the globally shared training dataset to train their local models and then use the globally shared scoring dataset to calculate $F_{i,t}^{p}$ in CB-DSL.

C. Evaluation and Comparison

In Fig. 1, when CB-DSL runs without $\mathfrak{D}_{tr}^{G}$, it cannot work properly in the non-i.i.d. setting. This is because CB-DSL hinges on single best worker selection which however may not hold the optimum model at all. Using $\mathfrak{D}_{sc}^{G}$ can slightly improve the learning performance of CB-DSL.When both a globally shared training dataset and scoring dataset are used as $\mathfrak{D}^{G} = \mathfrak{D}_{tr}^{G} \cup \mathfrak{D}_{sc}^{G}$, CB-DSL turns to outperform FedAvg. This is because $\mathfrak{D}_{tr}^{G}$ helps to relieve the local data heterogeneity issue by making the local datasets to become more i.i.d., which decreases the EMD between the data distributions on local workers and the population distribution as revealed by our model divergence analysis in Section V. Besides, the improvement on learning accuracy also indicates that our CB-DSL solutions have an increased chance to jump out of local optimum traps via the exploration-exploitation mechanism.

In Fig. 2, we provide the performance comparison in the presence of the Byzantine attack. It is obvious that even only one Byzantine attacker can fail FedAvg and CB-DSL without $\mathfrak{D}^{G}$. On the other hand, the CB-DSL with $\mathfrak{D}^{G}$ can effectively defend the Byzantine attack, because the globally shared dataset for scoring $\mathfrak{D}_{sc}^{G}$ can help identify and screen out the Byzantine attacker as explained in **Algorithm 1**.

In Fig. 3, we further evaluate the weight divergences effects under the non-i.i.d. setting. As the communication rounds increase, the weight divergences of CB-DSL with or without $\mathfrak{D}^{G}$ first increase and then flatten out after several communication rounds. The final steady-state weight divergence of the CB-DSL with $\mathfrak{D}^{G}$ is much less than that of the CB-DSL without $\mathfrak{D}^{G}$, as depicted by the gap between the two curves in Fig. 3. Such a nontrivial gap confirms the theoretical results of **Theorem 2**: (1) the model divergence will be enlarged as the communication rounds increase (this is because that the initial model divergence is iteratively amplified by β, as explained in *Remark 2*); (2) the use of $\mathfrak{D}^{G}$ can reduce the weight divergence (this is because that the use of $\mathfrak{D}^{G}$ decreases the EMD between the data distributions on local workers and the population distribution, as explained in *Remark 3*).

Note that only one local worker is selected and invited to send its model parameter to the PS in CB-DSL, while all workers need to send their model parameters to the PS in FedAvg. Therefore, the communication cost consumed in CB-DSL is only $\frac{1}{U}$ of that in FedAvg. In addition, we can see from Fig. 1 that our CB-DSL with $\mathfrak{D}^{G}$ uses fewer communication rounds than FedAvg to achieve the same learning accuracy. As a result, our CB-DSL is communication-efficient with less communication rounds and less communication overhead per round in practical applications.

VI. Conclusion

This work studies a novel communication-efficient and Byzantine-robust distributed swarm learning (CB-DSL) approach for edge IoT systems, as a holistic integration of the AI-enabled SGD and the BI-enabled PSO. We propose to introduce a globally shared dataset to overcome the major challenging issues in edge learning including: the partially observability of loss function in distributed learning problems, the non-i.i.d. local data issues, and the potential Byzantine attacks. We provide theoretical analysis of the convergence behavior of the proposed CB-DSL, which indicates that our method can achieve better learning performance than existing distributed learning methods. Further, we provide the model

[3] http://yann.lecun.com/exdb/mnist/

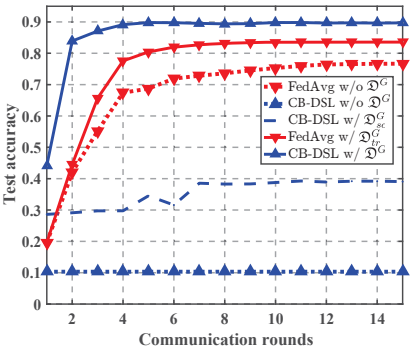

Fig. 1: The performance comparison under the non-i.i.d. setting.

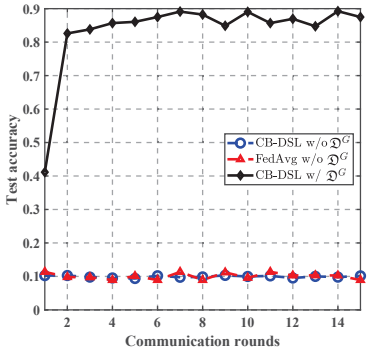

Fig. 2: The performance comparison with a Byzantine attacker.

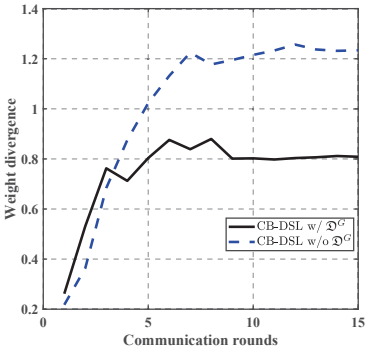

Fig. 3: The comparison of the weight divergences under the non-i.i.d. setting.

divergence evaluation of the proposed CB-DSL in the non-i.i.d. settings, which quantifies how a globally shared dataset can improve the learning performance of the CB-DSL. Simulation results verify that our proposed CB-DSL solution can improve learning performance in non-i.i.d. settings. Meanwhile, the communication saving by the CB-DSL inherits the advantage of the bio-inspired PSO techniques with much reduced communication cost than standard FedAvg.

ACKNOWLEDGMENTS

This work was partly supported by the Science and Technology Innovation Project of Xiongan (Grants #2022XACX0400), the National Science Foundation of the US (Grants #1939553, #2003211,#2128596, #2136202 and #2231209), and the Virginia Research Investment Fund (Commonwealth Cyber Initiative Grant #223996).

REFERENCES

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

[2] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "Joint optimization of communications and federated learning over the air," *IEEE Transactions on Wireless Communications*, vol. 21, no. 6, pp. 4434–4449, 2022.

[3] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 269–283, 2020.

[4] X. Fan and Y. Huo, "Cooperative secure transmission against collusive eavesdroppers in internet of things," *International Journal of Distributed Sensor Networks*, vol. 16, no. 6.

[5] ——, "Security analysis of cooperative jamming in internet of things with multiple eavesdroppers," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[6] 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.

[7] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik, "Sgd: General analysis and improved rates," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5200–5209.

[8] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[9] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[10] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[11] Z. Huo and H. Huang, "Asynchronous stochastic gradient descent with variance reduction for non-convex optimization," *arXiv preprint arXiv:1604.03584*, 2016.

[12] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "Bev-sgd: Best effort voting sgd against byzantine attacks for analog-aggregation-based federated learning over the air," *IEEE Internet of Things Journal*, vol. 9, no. 19, pp. 18 946–18 959, 2022.

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

[14] G. Zhou, P. Xu, Y. Wang, and Z. Tian, "Robust distributed learning against both distributional shifts and byzantine attacks," *arXiv preprint arXiv:2210.16682*, 2022.

[15] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "Best effort voting power control for byzantine-resilient federated learning over the air," in *2022 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2022, pp. 1–6.

[16] ——, "Communication-efficient federated learning through 1-bit compressive sensing and analog aggregation," in *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2021, pp. 1–6.

[17] Y. Wang, Z. Tian, X. Fan, Y. Huo, C. Nowzari, and K. Zeng, "Distributed swarm learning for internet of things at the edge: Where artificial intelligence meets biological intelligence," *arXiv preprint arXiv:2210.16705*, 2022.

[18] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "Joint optimization for federated learning over the air," in *2022 IEEE International Conference on Communications (ICC 2022)*. IEEE, 2022, pp. 1–6.

[19] F. Hussain, R. Hussain, S. A. Hassan, and E. Hossain, "Machine learning in IoT security: Current solutions and future challenges," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1686–1721, 2020.

[20] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "1-bit compressive sensing for efficient federated learning over the air," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2022.

[21] C. Selvaraj, R. S. Kumar, and M. Karnan, "A survey on application of bio-inspired algorithms," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 1, pp. 366–70, 2014.

[22] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, vol. 4. IEEE, 1995, pp. 1942–1948.

[23] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *MHS'95. Proceedings of the sixth international symposium on micro machine and human science*. Ieee, 1995, pp. 39–43.

[24] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F.-Y. Wang, "Generative adversarial networks: Introduction and outlook," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 4, pp. 588–598, 2017.

[25] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "Cb-dsl: Communication-efficient and byzantine-robust distributed swarm learning on non-iid data," *arXiv preprint arXiv:2208.05578*, 2022.

[26] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International journal of computer vision*, vol. 40, no. 2, pp. 99–121, 2000.