Communication Resources Limited Decentralized Learning with Privacy Guarantee through Over-the-Air Computation

Jing Qiao[1], Shikun Shen[1], Shuzhen Chen[1], Xiao Zhang[1*], Tian Lan[2], Xiuzhen Cheng[1], Dongxiao Yu[1*]
[1]School of Computer Science and Technology, Shandong University, China
[2]Department of Electrical and Computer Engineering, George Washington University, USA
{jingq,shikunshen,szchen}@mail.sdu.edu.cn,xiaozhang@sdu.edu.cn,tlan@gwu.edu,{xzcheng,dxyu}@sdu.edu.cn

ABSTRACT

In this paper, we propose a novel decentralized learning algorithm, namely *DLLR-OA*, for resource-constrained over-the-air computation with formal privacy guarantee. Theoretically, we characterize how the limited resources induced model-components selection error and compound communication errors jointly impact decentralized learning, making the iterates of *DLLR-OA* converge to a contraction region centered around a scaled version of the errors. In particular, the convergence rate of the *DLLR-OA* algorithm in the error-free case $O(\frac{1}{\sqrt{nT}})$ achieves the state-of-the-arts. Besides, we formulate a power control problem and decouple it into two sub-problems of transmitter and receiver to accelerate the convergence of the *DLLR-OA* algorithm. Furthermore, we provide quantitative privacy guarantee for the proposed over-the-air computation approach. Interestingly, we show that network noise can indeed enhance privacy of aggregated updates while over-the-air computation can further protect individual updates. Finally, the extensive experiments demonstrate that *DLLR-OA* performs well in the communication resources constrained setting. In particular, numerical results on CIFAR-10 dataset shows nearly 30% communication cost reduction over state-of-the-art baselines with comparable learning accuracy even in resource constrained settings.

CCS CONCEPTS

• **Computing methodologies** → **Distributed algorithms**.

KEYWORDS

decentralized learning, over-the-air computation, resource allocation, privacy-preserving

ACM Reference Format:
Jing Qiao[1], Shikun Shen[1], Shuzhen Chen[1], Xiao Zhang[1*], Tian Lan[2], Xiuzhen Cheng[1], Dongxiao Yu[1*]. 2023. Communication Resources Limited Decentralized Learning with Privacy Guarantee through Over-the-Air Computation. In *The Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc '23), October 23–26, 2023, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3565287.3610268

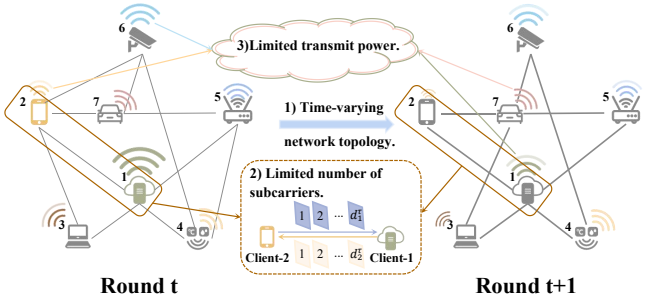


Figure 1: An illustration of key communication challenges in a decentralized learning framework: $\tau = t$ for the left and $\tau = t + 1$ for the right.

1 INTRODUCTION

Recently, distributed machine learning has attracted much research attention [14, 19] due to the explosive growth of data at network edge. Federated Learning (FL) [15] and fully decentralized learning [13] are two widely adopted distributed machine learning approaches to support intelligent data analytics and learning in edge computing environments. In particular, fully decentralized learning [13] is a dominating approach in areas such as the Internet of Vehicles [25], where clients communicate only with their neighbors thus forming an arbitrary (potentially time-varying) topology without relying on the coordination of a central server.

While wireless networks are often used in fully decentralized learning to support the high mobility of edge devices [24], it could easily become a performance bottleneck due to frequent information exchange among the computing clients and the resulting high communication overhead. Most existing works on communication-effective learning algorithms employ compression techniques, e.g., sparsification and quantization, to mitigate the communication overhead. Distributed stochastic gradient descent by exchanging sparse updates instead of dense updates was proposed in [1]. Related works also include Quantized SGD (QSGD) [2] enabling a tradeoff between compression and convergence speed, as well as algorithms [12] increasing local training sessions. However, these existing works either fail to provide a theoretical convergence guarantee, or ignore time-varying network conditions/topologies and resource constraints in wireless transmission.

To this end, over-the-air computation in decentralized learning (in contrast to the traditional "communication-then-aggregation" mechanism) has been proposed to utilize the superposition property

*Xiao Zhang and Dongxiao Yu are corresponding authors.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MobiHoc '23, October 23–26, 2023, Washington, DC, USA
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9926-5/23/10…$15.00
https://doi.org/10.1145/3565287.3610268

of the wireless channels to complete aggregation during transmission in the physical wireless channel [4]. As shown in Figure 1, over-the-air computation must account for wireless resource constraints (e.g., the number of subcarriers and available transmit power), as well as the dynamism of network conditions and time-varying topologies, in the learning algorithms. In this paper, we address a number of key challenges in over-the-air computation: (1) **model-components selection error.** In over-the-air computation, the number of subcarriers among each connection/link might be limited due to wireless bandwidth constraints. Therefore, only part of the model parameters can be transmitted to its neighbors in each step [27], leading to a model-components selection error that needs to be analyzed. (2) **Compound communication errors.** Since wireless channels are noisy, the received model parameters in over-the-air computation could suffer from compound communication errors. This requires novel solutions to optimize client device's individual transmit power, with the goal of maximizing resulting model accuracy in decentralized learning. (3) **Dynamic network conditions.** The high mobility of edge devices leads to time-varying network topologies. In addition, the constraints on communication resources are not fixed due to the unstable network. The above dynamically changing network conditions make the communication mode more complicated, which brings challenges to convergence analysis and optimization of decentralized learning. (4) **Privacy guarantee.** Finally, wireless channel noise and over-the-air computation mechanism also present a unique opportunity for providing privacy protection in distributed learning. Rigorously characterizing the privacy guarantee has not been considered.

This paper proposes a novel decentralized learning algorithm, namely *DLLR-OA*, for resource-constrained over-the-air computation with formal privacy guarantee. Firstly, with respect to wireless channel noise and limited number of subcarriers available for each wireless link, we theoretically characterize how model-components selection error and compound communication errors jointly impact the convergence of decentralized learning under the dynamic network conditions and time-varying topologies. In particular, we prove that the iterates of *DLLR-OA* would converge to a neighborhood of the scale of these two types of errors. Next, we formulate an optimization problem for minimizing the communication error (in terms of *MSE*) to accelerate the convergence of *DLLR-OA*. Decoupling into two sub-problems of transmitter and receiver, a power control problem is formulated and solved through a two-step operation consisting of scaling and recovery steps. Finally, we quantitatively analyze the privacy guarantee for the proposed over-the-air computation approach. To analyze the privacy-preserving mechanism of network noise, we leverage differential privacy (DP) technique, and solve the key problem on how to bound L_2-sensitivity in this technique through power constraints and inequality transformation. Intuitively, over-the-air computation can protect individual information because it completes aggregation during the communication process, but there is currently a lack of theoretical analysis, we here use the properties of the solutions to linear equations to give a formal mathematical support. Interestingly, we show that network noise can indeed enhance privacy of aggregated updates while over-the-air computation can further protect individual updates.

Our key contributions are summarized as follows:

- We propose a decentralized learning algorithm, namely *DLLR-OA*, for resource constrained over-the-air computation with formal privacy guarantee.
- The convergence of *DLLR-OA* is quantified under dynamic network conditions/topologies, model-components selection error due to limited bandwidth, and compound communication errors due to channel noise. We prove that the iterates of *DLLR-OA* would converge to a small neighborhood of the scale of these errors and at a rate of $O(\frac{1}{\sqrt{nT}})$ in the resource-unconstrained setting, suffering no loss in convergence speed compared with state-of-the-arts.
- We provide quantitative privacy guarantee for *DLLR-OA* by analyzing how the existence of channel noise enhances privacy in aggregation of neighboring information and how over-the-air aggregation protect individual updates from potential eavesdropping.
- We perform extensive experiments to evaluate the performance of *DLLR-OA*. Numerical results on CIFAR-10 dataset shows nearly 30% communication cost reduction over state-of-the-art baselines with comparable learning accuracy even in resource constrained settings.

2 RELATED WORK

In recent years, decentralized learning that relies on large-scale data and high-dimensional models often has extremely high demands on communication resources. Therefore, it is important to study communication-efficient decentralized learning to obtain higher performance with limited network resources. Such algorithms in the decentralized learning literature are based on compression, such as sparsification [1, 22] and quantization [2, 3], assuming lossless communication. In practice, however, communication is often lossy due to unstable networks and limited resources.

Classical model/gradient transmission often suffers from privacy leakage due to model inversion and reconstruction attacks [7, 9]. Differential privacy (DP) [5] methods can achieve a certain level of privacy protection within a given privacy budget by adding noise. Farokhi et al. [6] considered asynchronous collaborative algorithms for machine learning models with DP settings. Wei et al. [23] show that the proposed method NbAFL can satisfy DP by correctly tuning the artificial noise variance. Seif et al. [20] proposed a FL framework under the local differential privacy and showed that the superposition property of the simulation scheme is beneficial for privacy preservation.

Over-the-air computation exploits the superposition property of multi-access channel (MAC) to ensure that communication and aggregation can be done simultaneously, which facilitates enhanced communication efficiency and reduced training latency [4, 18, 21]. In addition, the paper [4] pointed out that the potential eavesdroppers can only access the aggregated updates instead of individual ones, which can protect private data.

Due to the advantages gradually shown by over-the-air computation, some works have started to investigate it in conjunction with various decentralized learning frameworks. Ozfatura et al. [17] focused on decentralized stochastic gradient descent (DSGD) taking into account the physical channel characteristics, without essential results on theoretical convergence. Shi et al. [21] proposed

an AirComp-based DSGT-VR algorithm in decentralized FL, where both precoding and decoding strategies at devices are developed to guarantee algorithm convergence. Unfortunately, this work is based on the assumption of sufficient available resources, without considering the poor network conditions. Michelusi [16] presented NCOTA-DGD to solve distributed machine learning problems over wirelessly-connected systems. But its assumption that the channels are noiseless and static cannot be applied to complex networks that are dynamically changing in reality.

In summary, under dynamic networks and constrained resources, decentralized learning via over-the-air computation with detailed theoretical convergence results and privacy analysis has not yet been well investigated.

3 PRELIMINARIES

3.1 Decentralized Learning Model

We consider the decentralized learning scenario with n clients $\mathcal{V} = \{1, \cdots, n\}$, $\mathcal{W} = (W_{ij})_{n\times n}$ is a doubly stochastic matrix, $W_{ij} > 0$ if client-i and client-j can communicate with each other. During the learning process, time is divided into synchronous rounds. In each round, client-i receives information aggregated over the air from all its neighbors, and then updates the model by performing local training using the aggregated neighbor information and its own local information.

The general learning problem is as follows:

$$\min_{\theta\in\mathbb{R}^d} f(\theta) := \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbb{E}_{\xi_i\sim\mathcal{D}_i}[F_i(\theta,\xi_i)]}_{:=f_i(\theta)} \tag{1}$$

In this paper, we consider a general dynamic scenario where the connections between clients can vary arbitrarily after each round. i.e., The neighboring set $N_i^t = \{j | W_{ij}^t > 0, j \in \mathcal{V}, j \neq i\}$ of client-i and weight matrix $\mathcal{W}^t = (W_{ij}^t)_{n\times n}$ of clients vary with the rounds.

3.2 Over-the-Air Aggregation

In the above decentralized learning scenario, we complete the process of client-i receiving information from all its neighbors client-j $\in N_i^t$ through over-the-air computation. That is, based on MISO communication, client-i receives aggregated information from all its neighbors over the wireless multiple access channel (MAC).

Specifically, each component of the parameters required by client-i is considered to be carried by one subcarrier of the channel. Thus, in round t of decentralized learning, the signal in subcarrier-k received by client-i can be expressed as:

$$y_i^t(k) = \sum_{j\in N_i^t} b_{ij}^t(k) h_{ij}^t(k) x_{ij}^t(k) + n_i^t(k) \tag{2}$$

when client-j sends message to client-i through subcarrier-k in round t, $b_{ij}^t(k)$ is the transmit power scaling factor, $h_{ij}^t(k)$ is the channel gain, $b_{ij}^t(k)x_{ij}^t(k)$ is the power of client-j when it transmits message to client-i through subcarrier-k, and $n_i^t(k)$ is the channel noise. In this paper, $x_{ij}^t(k)$ represents a component of the local model of client-j transmitted to client-i via subcarrier-k in round t. If the model-components of the corresponding coordinates of all neighboring client models are transmitted to client-i through subcarrier-k, the information received by client-i is already aggregated because of the superposition property of the channel.

In Table 1, we summarize the main notations in this paper.

Table 1: Notations and Descriptions

Notations	Descriptions
θ_i^t	The local model parameter of client-i in round t
η	The step length of algorithm *DLLR-OA*
ε	The budget of DP mechanism
δ	The slack variable of DP mechanism
$\mathcal{D}_i$	The local data set of client-i
T	The total round for algorithm *DLLR-OA*
t	The current round for algorithm *DLLR-OA*
$F_i(\theta_i^t, \xi_i^t)$	The loss of client-i on data sample ξ_i^t in round t
$y_i^t(k)$	The signal in subcarrier-k received by client-i in round t
$x_{ij}^t(k)$	The signal in subcarrier-k sent by client-j to client-i in round t
$b_{ij}^t(k)$	The transmit power scaling factor
$h_{ij}^t(k)$	The channel gain
$n_i^t(k)$	The channel noise and $n_i^t(k) \sim \mathcal{N}(0, \sigma^2)$
m_i^t	The mask generated by client-i based on its local model-components selection in round t
g_i^t	The gradient of client-i in round t, $g_i^t = \nabla F_i(\theta_i^t, \xi_i^t)$
$f_i(\theta_i^t)$	The loss function of client-i in round t, $f_i(\theta_i^t) = \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[F_i(\theta_i^t, \xi_i^t)]$
r_i^t	The model-components selection error of client-i in round t, $r_i^t = \sum_{j\in N_i^t} W_{ij}^t (C_i^t(\theta_j^t) - \theta_j^t)$
ϵ_i^t	The communication error of client-i in round t, $\epsilon_i^t = \hat{R}_i^t - \sum_{j\in N_i^t} W_{ij}^t C_i^t(\theta_j^t)$
r^t	$r^t = [r_1^t, \cdots, r_n^t]^T$
ϵ^t	$\epsilon^t = [\epsilon_1^t, \cdots, \epsilon_n^t]^T$
G^t	$G^t = [g_1^t, \cdots, g_n^t]^T$
H^t	$H^t = [\nabla f_1(\theta_1^t), \cdots, \nabla f_n(\theta_n^t)]^T$
Θ^t	$\Theta^t = [\theta_1^t, \cdots, \theta_n^t]^T \in \mathbb{R}^{n\times d}$

4 DECENTRALIZED LEARNING WITH LIMITED RESOURCES THROUGH OVER-THE-AIR COMPUTATION

4.1 DLLR-OA Algorithm

With constrained subcarriers, it is often difficult for an arbitrary client-i to obtain information about all components of its neighbor models. We consider that for client-i, in each round t, it expects to get as much model-components information as possible from its neighbors $j \in N_i^t$ according to its *model-components selection strategy*. The number of selected model-components depends on the number of subcarriers of the corresponding channel.

Definition 1. *(**model-components Selection**). Suppose we have a d-dimensional model $\theta \in \mathbb{R}^d$ and a certain strategy, $m \in \{0,1\}^d$ is a mask generated based on the strategy, we can pick some components of the current model parameters by $\theta \odot m$. And for the l-th component of $\theta \odot m$, we have*

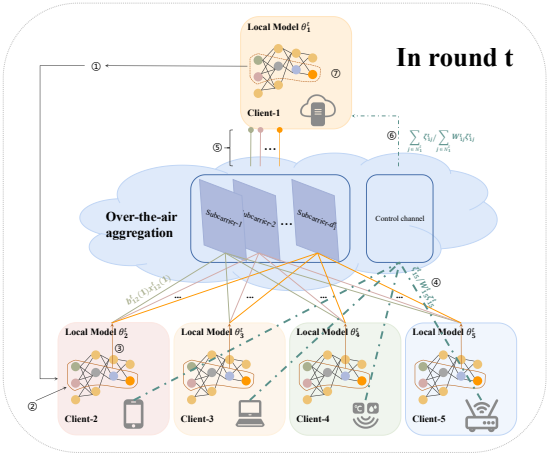


Figure 2: An overview of the process of the *DLLR-OA* algorithm, taking client-1 in Figure 1 as an example. ①: Generate a mask m_1^t and send it to all neighbors of client-1, ②: Select d_1^t model-components, ③: Transmit model-components information, ④: Transmit scaling factors information, ⑤: Aggregated information related to the model-components of all neighbors, ⑥: Aggregated information related to the scaling factors, ⑦: Use ⑥ to correct ⑤ for local model updates.

$$(\theta \odot m)_l = \begin{cases} (\theta)_l, & if\ (m)_l = 1 \\ 0, & if\ (m)_l = 0 \end{cases}$$

Specifically, we assume that for round t, the model-components selection strategy for client-i generates a mask m_i^t, where $m_i^t \in \{0,1\}^d$. Thus, the aggregated information that client-i expects to receive from neighboring clients can be expressed as $\sum_{j\in N_i^t} W_{ij}^t(\theta_j^t \odot m_i^t) = \sum_{j\in N_i^t} W_{ij}^t C_i^t(\theta_j^t)$, where W_{ij}^t is the weight between client-i and client-j at round t, $N_i^t = \{j | W_{ij}^t > 0, j \in \mathcal{V}, j \neq i\}$ is the neighboring set of client-i at round t, θ_i^t is the local model of client-i at round t and $C_i^t(\theta_j^t) = (\theta_j^t \odot m_i^t)$. We next elaborate the learning process in round t as shown in Figure 2:

- **Mask Generation and Transmission.** First, client-i finds the d_i^t coordinates that may contain more local information based on the local model-components selection strategy, then generates a mask m_i^t and transmits it to all neighboring clients (for all $j \in N_i^t$).
- **Receive and correct the Aggregate Sum of Neighboring Information** Then, every neighboring client-j will get $C_i^t(\theta_j^t) = \theta_j^t \odot m_i^t$ according to the mask m_i^t of client-i. Next, ideally, client-i will receive every non-zero component of $\sum_{j\in N_i^t} W_{ij}^t C_i^t(\theta_j^t)$ transmitted by client-$j \in N_i^t$. However, in practice, due to the transmit power limitation and the presence of channel noise, the model component information received by client-i carried by subcarrier-k can be represented as $(\widetilde{R}_i^t)_{I(k)} = y_i^t(k) = \sum_{j\in N_i^t} b_{ij}^t(k) h_{ij}^t(k) x_{ij}^t(k) + n_i^t(k)$. In this representation, $b_{ij}^t(k)$ is the transmit power scaling factor, which will be optimized in subsequent parts of this paper. $h_{ij}^t(k)$ is the time-varying channel gain. $x_{ij}^t(k)$ represents a component of the local model of client-j transmitted to client-i, $b_{ij}^t(k)x_{ij}^t(k)$ is the power of client-j when it transmits message to client-i through subcarrier-k, and $n_i^t(k)$ is the channel noise.
  After receiving $(\widetilde{R}_i^t)_{I(k)}$, client-$i$ can locally correct $(\widetilde{R}_i^t)_{I(k)}$ to obtain a good estimate $(\hat{R}_i^t)_{I(k)}$ of the corresponding component $(\sum_{j\in N_i^t} W_{ij}^t C_i^t(\theta_j^t))_{I(k)}$, and thus use $\hat{R}_i^t$ for subsequent local model updates.
- **Local Update** Finally, client-i updates its model using the neighbor information obtained through over-the-air aggregation and local information.

Algorithm 1 DLLR-OA

Input: The initial local model $\theta_i^0 \in \mathbb{R}^d$ $(i = 1, \cdots, n)$, the number of subcarriers d_i^t $(d_i^t \leq d, i = 1, \cdots, n, t = 0, 1, \cdots, T-1)$, step length η and client-i's model-components selection strategy $(i = 1, \cdots, n)$

Output: Local model θ_i $(i = 1, \cdots, n)$

for round $t = 0, 1, \cdots, T-1$ **do**
  // **Receiver Side:**
  for Receiver client-$i = 1, \cdots, n$ (In Parallel) **do**
    Generate a mask m_i^t based on the local model-components selection strategy;
    Send m_i^t to client-j, for all $j \in N_i^t$;
    Receive the sum of neighboring clients' information $\widetilde{R}_i^t$ through over-the-air aggregation;
    Correct $\widetilde{R}_i^t$ locally to get a good estimator $\hat{R}_i^t$ of $\sum_{j\in N_i^t} W_{ij}^t C_i^t(\theta_j^t)$;
    Update local model $\theta_i^{t+1} = \hat{R}_i^t + W_{ii}^t \theta_i^t - \eta \nabla F_i(\theta_i^t, \xi_i^t)$.
  end for
  // **Transmitter Side:**
  for Transmitter client-$j \in N_i^t$ $(i = 1, 2, \cdots, n)$ (In Parallel) **do**
    Receive m_i^t from client-i, where $\| m_i^t \|_1 = d_i^t$;
    Compute power allocation coefficients $b_{ij}^t(k)$ $(k = 1, 2, \cdots, d_i^t)$;
    Transmit $b_{ij}^t(k)x_{ij}^t(k)$ through subcarrier-k $(k = 1, 2, \cdots, d_i^t)$;
    Transmit information related to scaling factors.
  end for
end for

Algorithm 1 shows that client-i can receive signals carried by at most d_i^t subcarriers at round t, where d_i^t is determined by both the local model nature of client-i and the number of subcarriers.

Remark 1. *For client-i $(i \in \mathcal{V})$, the component coordinates and subcarriers are one-to-one mapping:*

$$(\widetilde{R}_i^t)_{I(k)} = \begin{cases} 0, & if\ (m_i^t)_{I(k)} = 0 \\ y_i^t(k), & if\ (m_i^t)_{I(k)} = 1 \end{cases}$$

4.2 Convergence Analysis of DLLR-OA

We next analyze the convergence rate of Algorithm 1. At first, we present the assumptions for this algorithm, which are widely used in decentralized learning [13].

Assumption 1. ***(Lipschitzian gradient)***. *Loss function $f_i(\cdot)$s are with Lipschitzian gradients. i.e., For $\forall \theta_i^{(1)}, \theta_i^{(2)} \in \mathbb{R}^d$, it holds that*

$$\| \nabla f_i(\theta_i^{(1)}) - \nabla f_i(\theta_i^{(2)}) \| \leq L \| \theta_i^{(1)} - \theta_i^{(2)} \|$$

Assumption 2. ***(Bounded variance)***. *The variance of the stochastic gradient is bounded as follows.*

$$\mathbb{E} \| \nabla F_i(\theta_i^t, \xi_i^t) - \nabla f_i(\theta_i^t) \|^2 \leq \sigma_1^2, \quad \forall i, \forall t$$
$$\mathbb{E} \| \nabla f_i(\theta) - \nabla f(\theta) \|^2 \leq \sigma_2^2, \quad \forall i$$

Assumption 3. ***(Symmetric double stochastic matrix)***. *In each round t, the communication matrix $\mathcal{W}^t$ is a real double stochastic matrix that satisfies $\mathcal{W}^t = (\mathcal{W}^t)^T$, $\mathcal{W}^t \mathbf{1}_n = \mathbf{1}_n$ and $\mathbf{1}_n^T \mathcal{W}^t = \mathbf{1}_n^T$.*

Assumption 4. ***(Spectral gap)***. *For any symmetric doubly stochastic matrix $\mathcal{W}^t$ above, we assume that $\rho_t = \max\{\lambda_2(\mathcal{W}^t), \lambda_n(\mathcal{W}^t)\} < 1$. And we write $\rho = \max_t \rho_t$.*

Based on the above assumptions, we can obtain the convergence result of *DLLR-OA* as Theorem 1.

Theorem 1. *Let*

$$D_1 = \frac{1-2L\eta}{2}, \quad D_2 = \Big(\frac{1}{2} - \frac{27nL^2\eta^2}{(1-\rho)^2\Big(1-\frac{54nL^2\eta^2}{(1-\rho)^2}\Big)}\Big),$$
$$D_3 = 1 - \frac{54nL^2\eta^2}{(1-\rho)^2}$$

If $\eta^2 \leq \min\{1, \frac{(1-\rho)^2}{108nL^2}\}$, we have the following result for Algorithm 1:

$$\begin{aligned}
&D_1 \cdot \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E} \| \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\theta_i^t) \|^2 + \\
&D_2 \cdot \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E} \| \nabla f(\bar{\theta}^t) - \frac{1}{2\eta D_2}\frac{\mathbf{1}_n^T}{n}(\epsilon^t + r^t) \|^2 \\
\leq &\frac{\mathbb{E}f(\bar{\theta}^0) - \mathbb{E}f(\bar{\theta}^*)}{\eta T} + \frac{3L^2}{2T(1-\rho^2)D_3}\mathbb{E} \| \Theta^0 \|^2 \\
&+ \frac{L^2\Big(\mathbb{E} \| \Theta^0 - \frac{\mathbf{1}_n\mathbf{1}_n^T}{n}\Theta^0 \|^2 - \mathbb{E} \| \Theta^* - \frac{\mathbf{1}_n\mathbf{1}_n^T}{n}\Theta^* \|^2\Big)}{2nTD_3} \\
&+ \Big(\frac{L\eta}{n} + \frac{3nL^2\eta^2}{(1-\rho^2)D_3}\Big)\sigma_1^2 + \frac{27nL^2\eta^2}{(1-\rho)^2 D_3}\sigma_2^2 \\
&+ \Big(\frac{9L^2}{(1-\rho)^2 D_3 T} + \frac{2L}{nT\eta} + \frac{1}{2D_2 nT\eta^2}\Big) \cdot \Big(\underbrace{\sum_{t=0}^{T-1} \mathbb{E} \| \epsilon^t \|^2}_{MSE} + \sum_{t=0}^{T-1} \Delta^{(t)}\Big)
\end{aligned}$$

Proof (Theorem 1). *Due to page limitation, we only outline the proof and state the important definitions and lemmas in the proof.*

Lemma 1 is an important property that is based on the assumptions on the network topology matrix $\mathcal{W}^t$. And its proof idea is similar to that of Lemma 5 in the paper [13]. With the help of Lemma 1, we can derive the subsequent Lemma 2 and 3 which are important for the convergence results.

Lemma 1. *Under Assumptions 3-4 above, we have*

$$\| \boldsymbol{e}_i^T \mathcal{W}^t \mathcal{W}^{t-1} \mathcal{W}^{t-2} \cdots \mathcal{W}^l - \frac{\mathbf{1}_n^T}{n} \| \leq \rho^{t-l+1}$$

Then, we elaborate the ***model-components selection error*** *in Definition 2:*

Definition 2. ***(model-components selection error)***. *In round t, model-components selection error can be written as follow:*

$$\begin{aligned}
\mathbb{E} \| r^t \|^2 &= \sum_{i=1}^{n} \mathbb{E} \| r_i^t \|^2 \\
&= \sum_{i=1}^{n} \mathbb{E} \| \sum_{j \in N_i^t} W_{ij}^t (C_i^t(\theta_j^t) - \theta_j^t) \|^2 \\
&= \Delta^{(t)}
\end{aligned}$$

And then we rewrite θ_i^{t+1} as:

$$\theta_i^{t+1} = \epsilon_i^t + r_i^t + \sum_{j=1}^{n} W_{ij}^t \theta_j^t - \eta \nabla F_i(\theta_i^t, \xi_i^t)$$

where $\epsilon_i^t = \hat{R}_i^t - \sum_{j \in N_i^t} W_{ij}^t C_i^t(\theta_j^t)$ is generated by communication aggregation. And $r_i^t = \sum_{j \in N_i^t} W_{ij}^t (C_i^t(\theta_j^t) - \theta_j^t)$ is caused by model-components selection. And this rewriting allows the impact of resource constraints to be represented visually in the subsequent results.

Next, Lemma 2 and Lemma 3 illustrate two bounds, using which the final convergence results can be directly derived.

Lemma 2. *If $\eta \in (0, \min\{1, \sqrt{\frac{(1-\rho)^2}{54nL^2}}\})$, under Assumption 1-4 above, we have*

$$\begin{aligned}
&\sum_{t=0}^{T-1} \mathbb{E} \| \Theta^t - \frac{\mathbf{1}_n\mathbf{1}_n^T}{n}\Theta^t \|^2 \\
\leq &\frac{\mathbb{E} \| \Theta^0 - \frac{\mathbf{1}_n\mathbf{1}_n^T}{n}\Theta^0 \|^2 - \mathbb{E} \| \theta^* - \frac{\mathbf{1}_n\mathbf{1}_n^T}{n}\theta^* \|^2}{\Big(1 - \frac{54nL^2\eta^2}{(1-\rho)^2}\Big)} \\
&+ \frac{3n}{(1-\rho^2)\Big(1 - \frac{54nL^2\eta^2}{(1-\rho)^2}\Big)}\mathbb{E} \| \Theta^0 \|^2 \\
&+ \frac{6n^2\eta^2\sigma_1^2}{1-\rho^2\Big(1 - \frac{54nL^2\eta^2}{(1-\rho)^2}\Big)}T + \frac{54n^2\eta^2\sigma_2^2}{(1-\rho)^2\Big(1 - \frac{54nL^2\eta^2}{(1-\rho)^2}\Big)}T \\
&+ \frac{54n^2\eta^2}{(1-\rho)^2\Big(1 - \frac{54nL^2\eta^2}{(1-\rho)^2}\Big)}\sum_{t=0}^{T-1} \mathbb{E} \| \nabla f(\bar{\theta}^t) \|^2 \\
&+ \frac{9n}{(1-\rho)^2\Big(1 - \frac{54nL^2\eta^2}{(1-\rho)^2}\Big)}\sum_{t=0}^{T-1} \mathbb{E} \| \epsilon^t + r^t \|^2
\end{aligned}$$

Lemma 3. *Under Assumption 1-4 above, we have*

$$\begin{aligned}
&\frac{\eta - 2L\eta^2}{2}\sum_{t=0}^{T-1} \mathbb{E} \| \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\theta_i^t) \|^2 + \frac{\eta}{2}\sum_{t=0}^{T-1} \mathbb{E} \| \nabla f(\bar{\theta}^t) \|^2 \\
\leq &\mathbb{E}f(\bar{\theta}^0) - \mathbb{E}f(\bar{\theta}^*) + \frac{L\eta^2\sigma_1^2 T}{n} + \sum_{t=0}^{T-1} \mathbb{E}\Big\langle \frac{\mathbf{1}_n^T}{n}(\epsilon^t + r^t), \nabla f(\bar{\theta}^t)\Big\rangle \\
&+ L\sum_{t=0}^{T-1} \mathbb{E} \| \frac{\mathbf{1}_n^T}{n}(\epsilon^t + r^t) \|^2 + \frac{L^2\eta}{2n}\sum_{t=0}^{T-1} \mathbb{E} \| \Theta^t - \frac{\mathbf{1}_n\mathbf{1}_n^T}{n}\Theta^t \|^2
\end{aligned}$$

Substituting Lemma 2 into Lemma 3, Theorem 1 can be further deduced. □

Theorem 1 characterizes the convergence rate of the average gradient of all local optimization variables θ_i^t and the gradient of the average local model $\bar{\theta}^t$ with the communication error and the model-components selection error. We choose an appropriate step length η in Theorem 1 to derive the following result.

Corollary 1. *Let* $\eta = \frac{1}{2L+\sigma_1\sqrt{T/n}}$. *If* $T \geq \frac{216n^2L^2}{\sigma_1^2(1-\rho)^2}$ *and* $T \geq \frac{9n^5L^4\left(\frac{\sigma_1^2}{1-\rho^2}+\frac{9\sigma_2^2}{(1-\rho)^2}\right)^2}{\sigma_1^6(\mathbb{E}f(\bar{\theta}^0)-\mathbb{E}f(\bar{\theta}^*)+L)^2}$, *we have*

$$D_2 \cdot \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\parallel \nabla f(\bar{\theta}^t) - \frac{1}{2\eta D_2}\frac{\mathbf{1}_n^T}{n}(\epsilon^t + r^t)\parallel^2$$

$$\leq Bound_1 + \Big(\frac{9L^2}{(1-\rho)^2 D_3 T} + \frac{4L^2}{nT} + \frac{2L\sigma_1}{n\sqrt{nT}} + \frac{1}{2D_2}\Big(\frac{2L}{\sqrt{nT}} + \frac{\sigma_1}{n}\Big)^2\Big) \cdot \Big(\underbrace{\sum_{t=0}^{T-1}\mathbb{E}\parallel \epsilon^t \parallel^2}_{MSE} + \sum_{t=0}^{T-1}\Delta^{(t)}\Big)$$

where the communication aggregation error term and the model-components selection error term are not included in $Bound_1$*:*

$$Bound_1 = \frac{2L}{T}\Big(\mathbb{E}f(\bar{\theta}^0) - \mathbb{E}f(\bar{\theta}^*)\Big) + \frac{3L^2}{2T(1-\rho^2)D_3}\mathbb{E}\parallel \Theta^0 \parallel^2 + \frac{L^2\Big(\mathbb{E}\parallel \Theta^0 - \frac{\mathbf{1}_n\mathbf{1}_n^T}{n}\Theta^0\parallel^2 - \mathbb{E}\parallel \Theta^* - \frac{\mathbf{1}_n\mathbf{1}_n^T}{n}\Theta^*\parallel^2\Big)}{2nTD_3} + \frac{2\sigma_1\Big(\mathbb{E}f(\bar{\theta}^0) - \mathbb{E}f(\bar{\theta}^*) + L\Big)}{\sqrt{nT}}.$$

In Corollary 1, the bound sharply decreases as the number of clients n and training round T increase. And it also shows that the communication error *MSE* caused by over-the-air aggregation may inhibit the convergence of the decentralized learning process. We can also know another fact that if there is no communication error and no model-components selection error, Corollary 1 indicates a convergence rate of $O(\frac{1}{\sqrt{nT}})$ when T is sufficiently large.

4.3 Transmit Power Allocation of DLLR-OA

From the above analysis, we find that the convergence result of Algorithm 1 may be affected by the communication error MSE. In this part, we will consider minimizing MSE by proper power allocation to speed up the convergence.

We assume that the channel state information is only available at the corresponding transmitter. i.e., only the client-j has the information of $\boldsymbol{h}_{ij}^t = [h_{ij}(1)^t, \cdots, h_{ij}^t(d_i^t)]^T$ in the process of sending signals from client-j to client-i.

Theoretically, if the transmit power of each client is not limited, we can set $b_{ij}^t(k) = \frac{W_{ij}^t}{h_{ij}^t(k)}$ to ensure that client-i receives an unbiased estimate of aggregated neighboring information. However, when the transmit power of each client is constrained, which is often a very common situation in real life, it is difficult for the client-i to receive an unbiased estimate of the transmitted signal.

In round t, all client-$j \in N_i^t$ optimize their local power allocation for transmitting the selected model-components over the d_i^t subcarriers to client-i, aiming to minimize the communication error so as to achieve a good estimation of $\sum_{j\in N_i^t} W_{ij}^t C_i^t(\theta_j^t)$ (or its scaled version).

Since the power needs to be re-allocated in each round, in order to simplify notation, we omit t when it is clear from the context in the following.

For the neighboring model-components information in the subcarrier-k ($k = 1, \cdots, d_i$) received by client-i, we introduce the model-components estimator coefficients $\{\alpha_i(k)\}_{k=1}^{d_i}$ (the model-components estimation error incurred by the lossy communication) to correct the aggregated information obtained. That is to say, $(\hat{R}_i^t)_{I(k)} = \alpha_i(k)\big(\sum_{j\in N_i} b_{ij}(k)h_{ij}(k)x_{ij}(k) + n_i(k)\big)$ is actually an estimator of $(\sum_{j\in N_i} W_{ij}C_i^t(\theta_j))_{I(k)} = \sum_{j\in N_i} W_{ij}x_{ij}(k)$. Then the aggregation error in every round can be denoted as:

$$MSE = \sum_{i=1}^{n}\Big(\sum_{k=1}^{d_i}\big(\sum_{j\in N_i}[\alpha_i(k)b_{ij}(k)h_{ij}(k) - W_{ij}]x_{ij}(k)\big)^2 + \sum_{k=1}^{d_i}\sigma^2\alpha_i^2(k)\Big)$$

An intuitive power allocation strategy is to solve the following optimization problem:

$$\mathbf{P1:}\quad \min_{\boldsymbol{\alpha},\boldsymbol{b}} \quad MSE$$
$$\text{s.t.}\quad \sum_{k=1}^{d_i}|b_{ij}(k)x_{ij}(k)|^2 \leq E_{ij}, \quad \forall i \in \mathcal{V}, j \in N_i$$
$$\alpha_i(k) \geq 0, \quad \forall i \in \mathcal{V}, j \in N_i, k \in \{1, \cdots, d_i\}$$
$$b_{ij}(k) \geq 0, \quad \forall i \in \mathcal{V}, k \in \{1, \cdots, d_i\}$$

Obviously, it is difficult to optimize $\boldsymbol{\alpha}, \boldsymbol{b}$ simultaneously to solve **P1**, and a feasible approach [27] is to optimize $\boldsymbol{\alpha}, \boldsymbol{b}$ alternately. i.e., first initialize $\boldsymbol{b}$, then find $\boldsymbol{\alpha}$ that minimizes *MSE* and satisfies the constraint, then use the obtained $\boldsymbol{\alpha}$ to find $\boldsymbol{b}$ that satisfies the constraint and minimizes *MSE*, and so on alternately. This method not only has a large computational overhead but also requires global information for each iteration to optimize $\boldsymbol{\alpha}, \boldsymbol{b}$ for each client, which is difficult to implement in practice.

Therefore, we focus on designing a sub-optimal solution for the power allocation and the setting of the model component estimation coefficients. In the process of transmitting signals to client-i, our objective is to minimize the communication aggregation error with the guarantee that $\alpha_i(k)\big(\sum_{j\in N_i} b_{ij}(k)h_{ij}(k)x_{ij}(k) + n_i(k)\big)$ is an unbiased estimate of $\sum_{j\in N_i} W_{ij}x_{ij}(k)$. Specifically, this optimization problem **P2** can be divided into two sub-problems:

In the sub-problem of transmitter-j. Unbiased estimation of the model-components information is often not possible due to the constrained transmit power. This makes it meaningful to achieve an unbiased estimation of a scaled version of the model-components information. Therefore the goal of the transmitter is to rationally allocate the available power to get an unbiased estimation of the scaled version.

P2 (Transmitter-j):

$$\max_{\{b_{ij}(k)\}} \quad \zeta_{ij}$$
$$\text{s.t.}\quad \zeta_{ij}W_{ij}x_{ij}(k) - b_{ij}(k)h_{ij}(k)x_{ij}(k) = 0, \quad \forall k \in \{1, \cdots, d_i\}$$
$$\sum_{k=1}^{d_i}|b_{ij}(k)x_{ij}(k)|^2 \leq E_{ij},$$

$$b_{ij}(k) \geq 0, \quad \forall k \in \{1, \cdots, d_i\}$$

The first constraint above makes it possible for the transmitter-j to transmit signals unbiased from the scaled versions of the corresponding product of weights and model-components. And the second and third are the power constraints. By using Karush-Kuhn-Tucker (KKT) conditions, the solution of the power allocation problem for transmitter-j can be derived as follows:

$$\zeta_{ij}^* = \sqrt{\frac{E_{ij}}{W_{ij}^2 \sum_{k=1}^{d_i} \frac{x_{ij}^2(k)}{h_{ij}^2(k)}}}, \quad b_{ij}^*(k) = \frac{\zeta_{ij}^* W_{ij}}{h_{ij}(k)} \tag{3}$$

In the sub-problem of receiver-i. model-components estimation coefficients $\{\alpha_i(k)\}_{k\in\{1,\cdots,d_i\}}$ need to be set appropriately to minimize communication errors and to achieve unbiased estimation of model component information. (The scaled version is appropriately corrected by the model component estimation coefficients to approximate the unscaled version)

P2 (Receiver-i):

$$\min_{\{\alpha_i(k)\}} \sum_{k=1}^{d_i} \Big(\sum_{j\in N_i} \big(\alpha_i(k)\zeta_{ij} W_{ij} x_{ij}(k) - W_{ij} x_{ij}(k)\big)\Big)^2 + \sum_{k=1}^{d_i} \sigma^2 \alpha_i^2(k)$$

$$\text{s.t.} \sum_{j\in N_i} \big(\alpha_i(k)\zeta_{ij} W_{ij} x_{ij}(k) - W_{ij} x_{ij}(k)\big) = 0, \quad \forall k \in \{1, \cdots, d_i\}$$

$$\alpha_i(k) \geq 0, \quad \forall k \in \{1, \cdots, d_i\}$$

The first constraint indicates that the scaled version of the corresponding product of weights and model-components with the adjustment of the model-components estimation coefficients is zero deviation from the unscaled version. And for given $\{\zeta_{ij}^*\}$, it holds

$$\frac{1}{\max\limits_{j\in N_i} \zeta_{ij}^*} \leq \alpha_i^*(k) = \frac{\sum\limits_{j\in N_i} W_{ij} x_{ij}(k)}{\sum\limits_{j\in N_i} \zeta_{ij}^* W_{ij} x_{ij}(k)} \leq \frac{1}{\min\limits_{j\in N_i} \zeta_{ij}^*} \tag{4}$$

We note that in the above, $x_{ij}(k)$ is not available at receiver-i. So it is wise to use one of the following two considerations as an approximation to $\alpha_i^*(k)$:

$$\alpha_i^* \simeq \alpha_i^\dagger = \frac{|N_i|}{\sum\limits_{j\in N_i} \zeta_{ij}^*} \tag{5}$$

$$\alpha_i^* \simeq \alpha_i^\ddagger = \frac{\sum\limits_{j\in N_i} W_{ij}}{\sum\limits_{j\in N_i} W_{ij}\zeta_{ij}^*} \tag{6}$$

where $\sum\limits_{j\in N_i} \zeta_{ij}^*$ or $\sum\limits_{j\in N_i} W_{ij}\zeta_{ij}^*$ can be transmitted to receiver-i through a control channel.

The most important feature of this resource allocation scheme is that it is considered separately by the transmitter side and the receiver side. During the communication, the target information is manipulated in two steps—scaling and recovery—in order to utilize the available resources sufficiently.

Remark 2. *From the above discussion, if the transmit power constraints are large enough ($E_{ij} \geq W_{ij}^2 \sum_{k=1}^{d_i} \frac{x_{ij}^2(k)}{h_{ij}^2(k)}, \forall i, j$), with scaling factor ζ_{ij} and coefficient $\alpha_i(k)$ equal to 1, we can recover $b_{ij}(k) = \frac{W_{ij}}{h_{ij}}$ to make each receiver obtain an unbiased estimate of the aggregated neighboring information. However, a potential possibility of our mechanism is able to reduce the variance of the network noise while obtaining an unbiased estimate by transmitting a multiple of $\frac{W_{ij}x_{ij}(k)}{h_{ij}}$ and going through a subsequent recovery operation at the receiver side.*

Remark 3. *By solving the sub-problem **P2**, we obtain the transmit power scaling factor $b_{ij}^*(k)$ as in eq. (3). One possible solution to solve **P1** mentioned above is to alternately optimize $\boldsymbol{\alpha}$ and $\boldsymbol{b}$, which may cause a computational bottleneck. Since we have already obtained $b_{ij}^*(k)$ by solving **P2**, can we directly substitute it back to problem **P1** to obtain the optimal $\alpha_i(k)$ thus avoiding a lot of alternating optimizations? If so, after a simple algebraic transformation, the optimal $\alpha_i(k)$ of the **P1** problem can be expressed as*

$$\alpha_i(k) = \frac{\sum\limits_{j\in N_i} W_{ij} x_{ij}(k)}{\sum\limits_{j\in N_i} \zeta_{ij}^* W_{ij} x_{ij}(k) + \frac{\sigma^2}{\sum\limits_{j\in N_i} \zeta_{ij}^* W_{ij} x_{ij}(k)}} \tag{7}$$

*However, eq. (7) shows that receiver-i has access to the model component information $x_{ij}(k)$ of its neighbors, which is not reasonable in practice. Therefore, a presumptuous attempt to solve **P1** is not feasible.*

*Moreover, theoretically, if the variance σ^2 of the channel noise is much smaller than $\sum\limits_{j\in N_i} \zeta_{ij}^* W_{ij} x_{ij}(k)$, eq. (7) can be considered equivalent to the $\alpha_i^*(k)$ obtained by solving the receiver sub-problem of **P2** as in eq. (4).*

Remark 4. *Due to the unavailability of individual information, it is also difficult to solve **P2** in practical scenarios, which motivates an approximate solution. The approximation of $\alpha_i^*(k)$ in eq. (4) is given by eq. (5) or (6), and the corresponding error can be bounded as follows, respectively:*

$$|\alpha_i^*(k) - \alpha_i^\dagger| \leq \max\Big\{\frac{1}{\min\limits_{j\in N_i} \zeta_{ij}^*} - \frac{|N_i|}{\sum\limits_{j\in N_i} \zeta_{ij}^*}, \frac{|N_i|}{\sum\limits_{j\in N_i} \zeta_{ij}^*} - \frac{1}{\max\limits_{j\in N_i} \zeta_{ij}^*}\Big\} \tag{8}$$

$$|\alpha_i^*(k) - \alpha_i^\ddagger| \leq \max\Big\{\frac{1}{\min\limits_{j\in N_i} \zeta_{ij}^*} - \frac{\sum\limits_{j\in N_i} W_{ij}}{\sum\limits_{j\in N_i} W_{ij}\zeta_{ij}^*}, \frac{\sum\limits_{j\in N_i} W_{ij}}{\sum\limits_{j\in N_i} W_{ij}\zeta_{ij}^*} - \frac{1}{\max\limits_{j\in N_i} \zeta_{ij}^*}\Big\} \tag{9}$$

4.4 Privacy Guarantee of DLLR-OA

In this part, we give an analysis on privacy preservation of *DLLR-OA*. First, we introduce some definitions of privacy protection.

Definition 3. *(**Privacy Preserving**). [26] A mechanism $\mathcal{M}$: $\mathcal{M}(X) \to Y$ is privacy preserving if the input X cannot be uniquely derived from the output Y.*

Definition 4. *(**(ε, δ)-DP**). Given a dataset with domain $\mathcal{D}$ and range $\mathcal{R}$, a randomized mechanism $\mathcal{M}$ preserves (ε, δ)-DP if for any two adjacent datasets $d, d' \in \mathcal{D}$ and any subset of outputs $\mathcal{S} \subseteq \mathcal{R}$ it holds that*

$$Pr(M(d) \in \mathcal{S}) \leq e^\varepsilon Pr(M(d') \in \mathcal{S}) + \delta,$$

where $\varepsilon \geq 0$ is a constant and δ is the probability of breaking this lower bound.

The L_2-sensitivity of the query function can be used to analyze DP, we elaborate it in Definition 5.

Definition 5. *(L_2-sensitivity). For a vector-valued function $f : \mathcal{D} \to \mathbb{R}^d$, the L_2-sensitivity of f is*

$$\Delta_2 f = \max_{d_1, d_2 \in \mathcal{D}} \| f(d_1) - f(d_2) \|_2,$$

where d_1 and d_2 differ in at most one element.

Lemma 4. *(Gaussian Mechanism). [5] Let $\varepsilon \in (0, 1)$ be arbitrary. For $c^2 > 2ln(1.25/\delta)$, the Gaussian Mechanism with parameter $\sigma \geq c\Delta_2 f/\varepsilon$ is (ε, δ)-differentially private.*

We consider that each client-i is honest but curious about its neighboring model-components information. Since over-the-air aggregation makes each client-i receive neighboring model-components information in the form of an aggregated sum, we first analyze the privacy performance of Algorithm 1 by considering this aggregated sum as a whole, which is given by Theorem 2.

Theorem 2. *In round t, Algorithm 1 satisfies $(\varepsilon_i^t(k), \delta)$-DP for aggregated neighboring information in any subcarrier-k received by client-i, where*

$$\varepsilon_i^t(k) = \frac{2h_{\max}^t \sum_{j \in N_i^t} \sqrt{E_{ij}^t}}{\sigma} \sqrt{2ln\frac{1.25}{\delta}}$$

and $h_{\max}^t = \max_{i,j,k} h_{ij}^t(k)$.

Proof (Theorem 2). *We firstly bound the L_2-sensitivity of $z_i^t(k) = \sum_{j \in N_i^t} b_{ij}^t(k) h_{ij}^t(k) x_{ij}^t(k)$. Consider two datasets $\mathcal{D}$ and $\mathcal{D}'$, L_2-sensitivity can be expressed as*

$$\begin{aligned}
&\Delta_2 f \\
&= \| z_i^t(k, \mathcal{D}) - z_i^t(k, \mathcal{D}') \| \\
&= \| \sum_{j \in N_i^t} b_{ij}^t(k, \mathcal{D}) h_{ij}^t(k) x_{ij}^t(k, \mathcal{D}) - \sum_{j \in N_i^t} b_{ij}^t(k, \mathcal{D}') h_{ij}^t(k) x_{ij}^t(k, \mathcal{D}') \|
\end{aligned}$$

Since $\sum_{k=1}^{d_i} |b_{ij}^t(k) x_{ij}^t(k)|^2 \leq E_{ij}^t$, we have $|b_{ij}^t(k) x_{ij}^t(k)| \leq \sqrt{E_{ij}^t}$. And $h_{\max}^t = \max_{i,j,k} h_{ij}^t(k)$, then L_2-sensitivity can be bounded as

$$\Delta_2 f \leq 2h_{\max}^t \sum_{j \in N_i^t} \sqrt{E_{ij}^t}$$

□

Further, if client-i makes aggregation of neighboring model-components available to it by some means, obtaining information about the model-components of one of its neighbors is also impossible, and we give the result in Theorem 3.

Theorem 3. *At round t, for an honest but curious receiver-i, if $|N_i^t| > 1$, Algorithm 1 can preserve the privacy of each neighboring model component $x_{ij}^t(k)$.*

Proof (Theorem 3). *If client-i has access to aggregation of neighboring model-components by some means, $\sum_{j \in N_i^t} b_{ij}^t(k) h_{ij}^t(k) x_{ij}^t(k) = \sum_{j \in N_i^t} \zeta_{ij}^t W_{ij}^t x_{ij}^t(k)$ is available for client-i.*
Under the assumption that $|N_i^t| > 1$, whether client-i receives the model-components estimation coefficient $\alpha_i^t(k)$ in the form of (5) or (6), client-i can know all $\zeta_{ij}^t(k)$ ($\forall j \in N_i^t$) if and only if $\zeta_{ij_1}^t(k) = \zeta_{ij_2}^t(k)$ ($\forall j_1, j_2 \in N_i^t$ and $j_1 \neq j_2$). At this point, $\sum_{j \in N_i^t} W_{ij}^t x_{ij}^t(k)$ is available for client-i. However, this single equation has $|N_i^t|$ unknowns. Hence, client-i can not have a unique solution for $x_{ij}^t(k)$ ($j \in N_i^t$) since the number of unknowns $|N_i^t|$ is greater than the number of equations, which is 1. □

5 EXPERIMENTS

In this part, we perform extensive experiments to evaluate our work. The details are shown as follows.

5.1 Experimental Setup

In our experiments, we train ResNet-18 [8] model on MNIST [11] and CIFAR-10 [10] datasets in different resource-constrained scenarios. We evaluate our work against the following baselines with sufficient available resources:

- **Local** is implemented by each client using its own data based on SGD algorithm, without any communication.
- **D-PSGD** [13] is based on the SGD algorithm for parallel training of all clients, considering neither constrained communication resources nor network noise.
- **D-PSGD (noise)** takes into account network noise compared to D-PSGD, but is still based on unconstrained communication resources.

For an arbitrary client-$i \in \mathcal{V}$ in round t, we consider the following three model-components selection strategies according to the number of subcarriers d_i^t of the corresponding channel:

- **Strategy-1:** Randomly select d_i^t coordinates.
- **Strategy-2:** Select the top-d_i^t coordinates corresponding to the L_2 parametrization of the model-components.
- **Strategy-3:** Select the top-d_i^t coordinates corresponding to the L_2 parametrization of the gradient components.

For all clients in round t, we can get decentralized learning systems with different restriction levels of subcarriers LS by setting different $d_i^t \leq d$ ($i \in \mathcal{V}$), where $\theta_i \in \mathbb{R}^d$. Take MNIST dataset in Table 2 as an example, ① $LS = 1.00$, ② $LS = 0.50$, ④ $LS = 0.10$ indicate that each client can receive a full model, 50% of model, and 10% of model information, respectively. And ③ $LS = 0.50$ means the clients receiving 75% of models, 50% of models and 25% of models information are each 1/3. We can compare ② $LS = 0.50$ and ③ $LS = 0.50$ to explore the impact of heterogeneous limited subcarriers on decentralized learning performance. As for the transmit power limit E_{ij}^t, we let $E_{ij}^t = \beta(\sum_{k=1}^{d_i^t} (\frac{W_{ij}^t x_{ij}^t(k)}{h_{ij}^t(k)})^2)$, where $\beta > 1$ indicates excess transmit power, $\beta = 1$ indicates proper transmit power, and $\beta < 1$ indicates insufficient transmit power. By adjusting different β values, we can set different levels of transmit power constraints. In particular, we satisfy the heterogeneity of the limited transmit power by adding a Gaussian noise for each β in E_{ij}^t. Without loss of generality, we take the variance of the channel noise as $\sigma^2 = 0.0001$ and $\{h_{ij}(k)\}$ are independent and identically distributed Rayleigh random variables with mean 1. In all our experiments, the number of clients is 12, the batch size is 128, the number of local training epochs is 5 and the learning rate is 0.001.

Table 2: Performance comparison of decentralized learning with different restriction levels of subcarriers based on different model-components selection strategies.

Dataset	Restriction level of subcarriers	Strategy-1		Strategy-2		Strategy-3	
		Acc. (%)	Average Comm.cost (*MB*)	Acc. (%)	Average Comm.cost (*MB*)	Acc. (%)	Average Comm.cost (*MB*)
MNIST	Local	98.46	-	98.46	-	98.46	-
	D-PSGD LS=1.00	99.50	2836.55	99.50	2836.55	99.50	2836.55
	D-PSGD (noise) LS=1.00	99.49	2836.55	99.49	2836.55	99.49	2836.55
	① LS=1.00	99.49	2836.55	99.49	2836.55	99.49	2836.55
	② LS=0.50	99.42	1404.27	99.44	2804.52	99.51	2324.74
	③ LS=0.50	99.36	1416.05	99.03	2761.47	99.48	2322.40
	④ LS=0.10	99.58	263.74	98.49	2115.99	99.10	229.92
CIFAR-10	Local	64.03	-	64.03	-	64.03	-
	D-PSGD LS=1.00	83.13	2830.47	83.13	2830.47	83.13	2830.47
	D-PSGD (noise) LS=1.00	83.70	2830.47	83.70	2830.47	83.70	2830.47
	① LS=1.00	83.24	2830.47	83.24	2830.47	83.24	2830.47
	② LS=0.60	80.19	1660.75	82.26	2821.91	82.25	2045.41
	③ LS=0.60	79.74	1659.45	79.82	2816.92	82.25	2035.29
	④ LS=0.30	75.39	809.32	65.01	2744.09	81.56	1100.48

All the experiments are implemented in PyTorch 1.11.0, Python 3.8, Cuda 11.3. And we run them on a Cloud Server with AMD EPYC 7642 48-core processors and total 4 RTX 3090 GPUs in Ubuntu 20.04.

5.2 Numerical Results

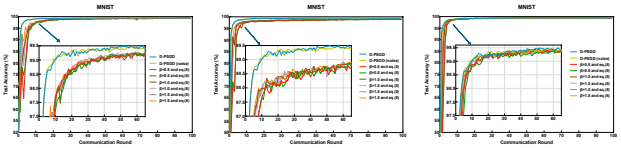
We evaluate our work using metrics: training loss, test accuracy, average communication cost and communication rounds. In particular, we explore the effects of the constraints on subcarries and transmit power separately.

Impact of limited subcarriers. On the MNIST and CIFAR-10 datasets, we compare the impact of limited subcarriers on decentralized learning under three different model-components selection strategies within 100 communication rounds.

As shown in Table 2, in general, the smaller the average number of subcarriers, the lower the test accuracy. The results show that the test accuracy of decentralized learning in resource-constrained situations outperforms that of fully local training method Local. For example, on CIFAR-10, the average test accuracy of ④ LS = 0.3 under the three strategies improved by 9.96% over that of Local. Furthermore, when the average restriction levels of subcarriers are same, the test accuracy of decentralized learning is lower with heterogeneous subcarrier restrictions (d_i^t may be different for varying i). Take ② and ③ on CIFAR-10 with Strategy-2 as an example, the case ③ with higher heterogeneity is 2.44% less accurate than the test results of ②. Further, the results show that transmitting the partial model rather than the full model determined by a proper model-components selection strategy is a communication-efficient mechanism. On MNIST, this mechanism achieves an accuracy comparable to traditional D-PSGD on the basis of reducing the communication cost by 91.89%. On CIFAR-10, it reduces the communication cost by 61.12%, bringing only 1.57% accuracy reduction.

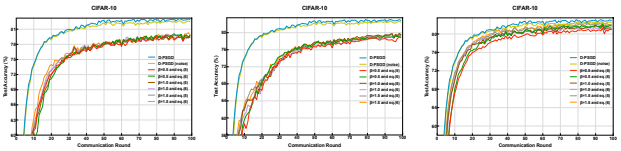
Impact of limited transmit power. On MNIST and CIFAR-10, we compare the impact of limited transmit power on decentralized learning under three different model-components selection strategies within 100 communication rounds. Note that here the number of subcarriers is also restricted, and the restriction levels correspond to ② $LS = 0.5$ for MNIST and ② $LS = 0.6$ for CIFAR-10 in Table 2.

As shown in Figure 3-4, proper transmit power ($\beta = 1.0$) or excess

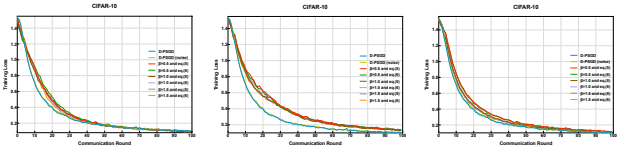


(a) model-components Selection Strategy-1 (b) model-components Selection Strategy-2 (c) model-components Selection Strategy-3

Figure 3: Test accuracy comparison of decentralized learning with different power limits on MNIST.



(a) model-components Selection Strategy-1 (b) model-components Selection Strategy-2 (c) model-components Selection Strategy-3



(d) model-components Selection Strategy-1 (e) model-components Selection Strategy-2 (f) model-components Selection Strategy-3

Figure 4: Performance comparison of decentralized learning with different power limits on CIFAR-10, with test accuracy on the top and training loss on the bottom.

transmit power ($\beta = 1.5$) leads to a faster convergence rate and a higher test accuracy than insufficient transmit power ($\beta = 0.5$). In particular, under model-components selection strategy-3, the excess transmit power ($\beta = 1.5$) enables decentralized learning with the limited number of subcarriers to reduce the communication cost by 29.04%, compared with D-PSGD and D-PSGD (noise) on CIFAR-10. However the accuracy of them is about the same. This result is due to the fact that during the communication process, the SNR is improved by amplifying the signal, thereby reducing

the impact of noise after subsequent recovery operations at the receiver side.

6 CONCLUSION

In this paper, we proposed the *DLLR-OA* algorithm integrating the communication resources allocation and privacy guarantee. Theoretically, we characterized the inhibition of the model-components selection error and compound communication errors caused by communication resources constraints on the convergence of decentralized learning. And we accelerated the convergence by designing an efficient resource allocation scheme. Moreover, we provided quantitative privacy guarantee with the help of differential privacy techniques and over-the-air computation mechanism. To further evaluate our work, we conducted sufficient experiments to show the possibility to achieve a high accuracy under communication resources constrained settings.

ACKNOWLEDGMENTS

This work was supported in part by National Science Fund for Excellent Young Scholars of China under Grant 62122042, in part by Major Basic Research Program of Shandong Provincial Natural Science Foundation under Grant ZR2022ZD02, in part by the Fundamental Research Funds for the Central Universities, in part by the National Natural Science Foundation of China under Grant 62202273, in part by Shandong Provincial Natural Science Foundation of China under Grant ZR2021QF044.

REFERENCES

[1] Alham Fikri Aji and Kenneth Heafield. 2017. Sparse Communication for Distributed Gradient Descent. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, 440–445. https://doi.org/10.18653/v1/d17-1045

[2] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. 2017. QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 1709–1720. https://proceedings.neurips.cc/paper/2017/hash/6c340f25839e6acdc73414517203f5f0-Abstract.html

[3] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. 2018. SIGNSGD: Compressed Optimisation for Non-Convex Problems. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 559–568. http://proceedings.mlr.press/v80/bernstein18a.html

[4] Xiaowen Cao, Zhonghao Lyu, Guangxu Zhu, Jie Xu, Lexi Xu, and Shuguang Cui. 2022. An Overview on Over-the-Air Federated Edge Learning. *CoRR* abs/2208.05643 (2022). https://doi.org/10.48550/arXiv.2208.05643 arXiv:2208.05643

[5] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.* 9, 3-4 (2014), 211–407. https://doi.org/10.1561/0400000042

[6] Farhad Farokhi, Nan Wu, David B. Smith, and Mohamed Ali Kâafar. 2021. The Cost of Privacy in Asynchronous Differentially-Private Machine Learning. *IEEE Trans. Inf. Forensics Secur.* 16 (2021), 2118–2129. https://doi.org/10.1109/TIFS.2021.3050603

[7] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, October 12-16, 2015*, Indrajit Ray, Ninghui Li, and Christopher Kruegel (Eds.). ACM, 1322–1333. https://doi.org/10.1145/2810103.2813677

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90

[9] Briland Hitaj, Giuseppe Ateniese, and Fernando Pérez-Cruz. 2017. Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017*, Bhavani Thuraisingham, David Evans, Tal Malkin, and Dongyan Xu (Eds.). ACM, 603–618. https://doi.org/10.1145/3133956.3134012

[10] Alex Krizhevsky. 2009. Learning Multiple Layers of Features from Tiny Images.

[11] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324. https://doi.org/10.1109/5.726791

[12] Xiang Li, Wenhao Yang, Shusen Wang, and Zhihua Zhang. 2019. Communication Efficient Decentralized Training with Multiple Local Updates. *CoRR* abs/1910.09126 (2019). arXiv:1910.09126 http://arxiv.org/abs/1910.09126

[13] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. 2017. Can Decentralized Algorithms Outperform Centralized Algorithms? A Case Study for Decentralized Parallel Stochastic Gradient Descent. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/f7552659f31040afeb61cb7133e4e6d-Paper.pdf

[14] Qin Liu, Liqiong Chen, Hongbo Jiang, Jie Wu, Tian Wang, Tao Peng, and Guojun Wang. 2022. A collaborative deep learning microservice for backdoor defenses in Industrial IoT networks. *Ad Hoc Networks* 124 (2022), 102727. https://doi.org/10.1016/j.adhoc.2021.102727

[15] H. Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. 2016. Federated Learning of Deep Networks using Model Averaging. *CoRR* abs/1602.05629 (2016). arXiv:1602.05629 http://arxiv.org/abs/1602.05629

[16] Nicolò Michelusi. 2022. Decentralized Federated Learning via Non-Coherent Over-the-Air Consensus. *CoRR* abs/2210.15806 (2022). https://doi.org/10.48550/arXiv.2210.15806 arXiv:2210.15806

[17] Emre Ozfatura, Stefano Rini, and Deniz Gündüz. 2020. Decentralized SGD with Over-the-Air Computation. In *IEEE Global Communications Conference, GLOBECOM 2020, Virtual Event, Taiwan, December 7-11, 2020*. IEEE, 1–6. https://doi.org/10.1109/GLOBECOM42002.2020.9322286

[18] E. Ozfatura, Stefano Rini, and D. Gündüz. 2020. Decentralized SGD with Over-the-Air Computation. In *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*. 1–6. https://doi.org/10.1109/GLOBECOM42002.2020.9322286

[19] Zhiqiang Pan, Fei Cai, Wanyu Chen, Chonghao Chen, and Honghui Chen. 2022. Collaborative Graph Learning for Session-based Recommendation. *ACM Trans. Inf. Syst.* 40, 4 (2022), 72:1–72:26. https://doi.org/10.1145/3490479

[20] Mohamed Seif, Ravi Tandon, and Ming Li. 2020. Wireless Federated Learning with Local Differential Privacy. In *IEEE International Symposium on Information Theory, ISIT 2020, Los Angeles, CA, USA, June 21-26, 2020*. IEEE, 2604–2609. https://doi.org/10.1109/ISIT44484.2020.9174426

[21] Yandong Shi, Yong Zhou, and Yuanming Shi. 2021. Over-the-Air Decentralized Federated Learning. In *IEEE International Symposium on Information Theory, ISIT 2021, Melbourne, Australia, July 12-20, 2021*. IEEE, 455–460. https://doi.org/10.1109/ISIT45174.2021.9517780

[22] Sebastian U. Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. 2018. Sparsified SGD with Memory. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). 4452–4463. https://proceedings.neurips.cc/paper/2018/hash/b440509a0106086a67bc2ea9df0a1dab-Abstract.html

[23] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H. Yang, Farhad Farokhi, Shi Jin, Tony Q. S. Quek, and H. Vincent Poor. 2020. Federated Learning With Differential Privacy: Algorithms and Performance Analysis. *IEEE Trans. Inf. Forensics Secur.* 15 (2020), 3454–3469. https://doi.org/10.1109/TIFS.2020.2988575

[24] Yu Ye. 2020. *Study on Decentralized Machine Learning and Applications to Wireless Caching Networks*. Ph. D. Dissertation. Royal Institute of Technology, Stockholm, Sweden. https://nbn-resolving.org/urn:nbn:se:kth:diva-276455

[25] Wei Yue, XiangJun Guo, Zhongchang Liu, Liyuan Wang, and Cunming Zou. 2022. Decentralized Robust Control for Internet of Connected Vehicles against Cyber-attacks. In *13th Asian Control Conference, ASCC 2022, Jeju, Korea, May 4-7, 2022*. IEEE, 409–414. https://doi.org/10.23919/ASCC56756.2022.9828199

[26] Chunlei Zhang, Muaz Ahmad, and Yongqiang Wang. 2019. ADMM Based Privacy-Preserving Decentralized Optimization. *IEEE Trans. Inf. Forensics Secur.* 14, 3 (2019), 565–580. https://doi.org/10.1109/TIFS.2018.2855169

[27] Junshan Zhang, Na Li, and Mehmet Dedeoglu. 2021. Federated Learning over Wireless Networks: A Band-limited Coordinated Descent Approach. In *40th IEEE Conference on Computer Communications, INFOCOM 2021, Vancouver, BC, Canada, May 10-13, 2021*. IEEE, 1–10. https://doi.org/10.1109/INFOCOM42981.2021.9488818