DReS-FL: Dropout-Resilient Secure Federated Learning for Non-IID Clients via Secret Data Sharing

Anonymous Author(s)
Affiliation
Address
email

Abstract

Federated learning (FL) strives to enable collaborative training of machine learning models without centrally collecting clients' private data. Different from centralized training, the local datasets across clients in FL are non-independent and identically distributed (non-IID). In addition, the data-owning clients may drop out of the training process arbitrarily. These characteristics will significantly degrade the training performance. This paper proposes a *Dropout-Resilient Secure Federated Learning* (DReS-FL) framework based on Lagrange coded computing (LCC) to tackle both the non-IID and dropout problems. The key idea is to utilize Lagrange coding to secretly share the private datasets among clients so that the effects of non-IID distribution and client dropouts can be compensated during local gradient computations. To provide a strict privacy guarantee for local datasets and correctly decode the gradient at the server, the gradient has to be a polynomial function in a finite field, and thus we construct *polynomial integer neural networks* (PINNs) to enable our framework. Theoretical analysis shows that DReS-FL is resilient to client dropouts and provides privacy protection for the local datasets. Furthermore, we experimentally demonstrate that DReS-FL consistently leads to significant performance gains over baseline methods.

1 Introduction

Federated learning (FL) [1] is a machine learning framework in which a central server coordinates a large number of clients to collaboratively train a shared model. The key idea of FL is to train the model locally by individual clients and aggregate updates globally by the server. The main target is to provide privacy protection for clients' local samples and solve the "data islands" problem. However, it has been shown recently that local models may reveal substantial information about the local datasets, and the private training data can be reconstructed through model inversion attacks [2, 3, 4]. Besides, as local data are typically non-independent and identically distributed (non-IID), the model divergence during the local update may lead to unstable and slow convergence [5, 6, 7]. With many clients involved in the training, some of the clients could drop out of the training process unexpectedly (due to poor connectivity, battery level, etc), and it will cause detrimental model performance [8]. Thus, effective mechanisms are needed to tackle the non-IID data distribution and client dropouts, while preserving the privacy of local datasets, which motivates this work.

To prevent information leakage from the local models, *secure aggregation* protocols [9, 10, 11, 12, 13, 14] have been developed to allow for global aggregation without revealing the parameters of clients' models. Even if some clients may drop out, these protocols can still recover the aggregated results of

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

the surviving clients. However, their training performance may degrade significantly in the non-IID setting due to the aggregation bias. To alleviate the non-IID problem, existing methods typically follow *algorithm-based approaches* [5, 15, 16, 17, 18] and add regularization terms to mitigate the model divergence. However, these methods are not dropout-resilient evidenced by the empirical results in [19]. This can be explained by the greatly varying data distributions among different rounds. Another fold of strategy for dealing with the non-IID problem is *data-centric approach* [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30], which generates extra training samples to construct a more balanced data distribution for each client. The common practices are to share the synthesized samples [22, 23, 24, 25] or GAN-based augmented data [26, 27, 28, 29, 30]. However, these methods may leak private information about local datasets and violate the privacy criterion in FL.

Contributions. In this work, we develop a ***D**ropout-**Re**silient **S**ecure **F**ederated **L**earning* (DReS-FL) framework to address the above problems via Lagrange coded computing (LCC) [31]. The key idea of LCC is to encode the datasets using Lagrange polynomials, in order to create computational redundancy across the workers in a privacy-preserving way. At the beginning of the training, the clients secretly share their private datasets with each other via Lagrange coding. This allows clients to access an encoded version of the *global dataset*[1] for local gradient computations while guaranteeing privacy in an information-theoretic sense. After collecting the computation results from a certain number of surviving clients, the server performs polynomial interpolation to decode the gradient, and thus it is resilient to client dropouts. As the local gradients are computed on the mini-batches uniformly sampled from the global dataset, the server obtains *global gradient*[2] after decoding. Therefore, the training process in DReS-FL is made equivalent to the centralized training and eliminates the non-IID problem. To provide a privacy guarantee for the local datasets and correctly decode the gradient at the server, the gradient has to be a polynomial function in a finite field, which is a main design challenge of DreS-FL. Our main contributions are summarized as follows:

- The proposed DReS-FL framework provides a unified approach to tackle two critical problems of FL, namely, *non-IID data distribution* and *client dropouts*. Meanwhile, it maintains privacy and security guarantees such that no information about local datasets can be leaked beyond the global model parameters.
- We construct *polynomial integer neural networks* (PINNs) to ensure that the gradient is a polynomial, so that cryptographic primitives can be applied for secure computation. A PINN consists of affine transformation layers with parameters constrained in an integer set, and it adopts the quadratic function as the activation function. The convergence analysis of DReS-FL with PINNs is also provided.
- We conduct extensive experiments on FL benchmark datasets to demonstrate the effectiveness of DReS-FL. It is shown that DReS-FL outperforms baseline methods under the setting where local datasets are heterogeneous and clients may drop out of the training process arbitrarily.

2 Related Works

Secure aggregation. Secure model aggregation [9, 10, 11, 12, 13, 14] is a key component of FL that protects the privacy of each client's model while allowing their global aggregation amidst possible user dropouts. Existing protocols essentially rely on two main principles, including a pairwise random-seed agreement for mask cancellation and secret sharing of the random seeds to construct the dropped masks [9, 10, 11, 12, 13, 14]. However, these approaches may suffer from severe performance degradation in non-IID FL, since the surviving clients in each round vary greatly, and thus the local gradients are biased towards different data distributions.

[1] In the context of this paper, we use the global dataset to denote the concatenation of all the clients' datasets.

[2] The global gradient corresponds to the stochastic gradient computed from mini-batches that are uniformly sampled from the global dataset. For simplicity, we consider that clients only perform one local stochastic gradient descent (SGD) step in each communication round. Note that the proposed DReS-FL framework, as stated in Remark 2, can be extended to more general cases in which clients can run multiple local SGD steps.

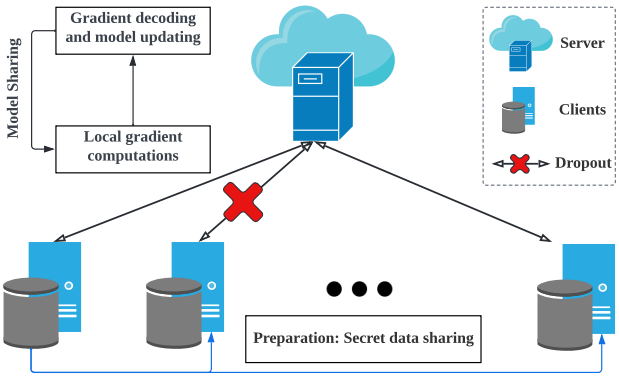


Figure 1: The DReS-FL system model. At the beginning of training, the clients secretly share the local datasets with each other. Then, the model parameters are iteratively trained by (1) local gradient computations and (2) gradient decoding and model updating until convergence.

Non-IID data and client dropouts. Training with heterogeneous data is a unique challenge for FL [1], which significantly affects the convergence performance [8]. The client dropouts exacerbate the non-IID problem as the data distributions among different rounds could vary greatly. Many algorithm-based methods [5, 15, 16, 17, 18] attempt to mitigate the clients' model divergence, but these methods cannot solve the essence of the non-IID problem due to the intrinsic difference between minimizing the local empirical loss and minimizing the global empirical loss. Another line of work adopts data-centric methods [26, 27, 28, 29, 30] to modify the local distributions. Ideally, a perfect data sharing mechanism should achieve that the local datasets have the same distribution as the global dataset while maintaining the privacy guarantee. Common practices include sharing raw datasets [20, 21], synthesized samples, [22, 23, 24, 25] or augmented data [26, 27, 28, 29, 30]. However, these works cannot fully preserve local data privacy in an information-theoretic sense [32]. A special data-centric method is the secret coding scheme, which has been widely utilized in homomorphic encryption (HE) [33, 34, 35, 36, 37, 38, 39] and multiparty computation (MPC) techniques [40, 41, 42]. This coding scheme allows computations to be performed on encrypted data and has been used for privacy-preserving machine learning [35, 39, 42]. However, the HE methods often suffer from time-consuming cryptographic tools, and MPC techniques are difficult to generalize such primitives to a large number of clients. Recently, distributed secure machine learning frameworks [43, 44] have been proposed for logistic regression problems. They apply Lagrange coding for secret data sharing and approximate the Sigmoid function by a polynomial function. This paper proposes DReS-FL to further extend these works to train deep neural networks in the FL setting.

3 System Model

We consider a federated learning framework as shown in Fig. 1 that consists of one central server and N data-owning clients. Each client $i \in [N]$ holds a local dataset $(\mathbf{X}_i, \mathbf{Y}_i)$ of size m_i, where $\mathbf{X}_i \in \mathbb{R}^{m_i \times d_x}$ represents the set of input features of dimension d_x and $\mathbf{Y}_i \in \mathbb{R}^{m_i \times d_y}$ corresponds to the output vector of dimension d_y. Accordingly, the size of the global dataset $(\mathbf{X}, \mathbf{Y})$ which concatenates all local datasets $(\mathbf{X}_i, \mathbf{Y}_i), \forall i \in [N]$ is denoted as $m \triangleq \sum_{i=1}^{N} m_i$. The clients aim to jointly train a neural network based on their local datasets without sharing private data samples. Particularly, the gradients are computed locally and aggregated globally. However, the local data may be highly heterogeneous, and the clients may drop out at any time unexpectedly, which makes the training process unstable. Our goal is to improve the convergence performance by secret data sharing while preserving the privacy of local datasets.

Threat model and privacy requirements. We consider a threat model where the clients are *honest-but-curious*. In particular, clients follow the protocol honestly but may collude amongst themselves to learn additional information. Besides, we assume that the server is also honest-but-curious, but does

not collude with any other clients. To avoid leaking private information from the shared samples and protect the local updates against model inversion attacks [2, 3, 4], we impose two privacy requirements in the training process. First, the clients learn nothing about the private datasets of others from the shared samples even if up to T clients collude. Second, the server cannot infer the private datasets from the local gradients uploaded from clients beyond the aggregated model.

Lagrange coded computing. The LCC framework considers a scenario involving computations over massive datasets stored distributedly across multiple clients [31]. The key idea is to encode the data using Lagrange polynomial for redundant distributed computing, which fits nicely with federated learning due to its dropout-resiliency and privacy guarantees. DReS-FL applies LCC to secretly share the private datasets among clients for local gradient computations, and the global gradient can be decoded by the server for model updating. To provide a strong privacy guarantee for the datasets and correctly decode the gradient at the server, the gradient should be a polynomial function in a finite field. However, existing neural networks cannot satisfy this requirement, since the datasets are in the real field and the gradients are not polynomial functions due to the non-linear operations.

Polynomial Integer Neural Networks. We define a class of polynomial integer neural networks (PINNs) to ensure that the gradient is a polynomial function in a finite field $\mathbb{F}_p$ with a prime number p. First, we transform the dataset $(\mathbf{X}, \mathbf{Y})$ from the real domain to the finite domain $(\overline{\mathbf{X}}, \overline{\mathbf{Y}})$. Besides, a PINN consists of affine transformation layers (e.g., fully connected layers and convolutional layers) and utilizes the quadratic function as the activation function. The model parameters of PINNs are defined in the integer set $\mathbb{Z}_p \triangleq \{-\lfloor \frac{p+1}{2} \rfloor, \ldots, \lfloor \frac{p-1}{2} \rfloor\}$. Given a feed-forward function $\boldsymbol{f}(\overline{\mathbf{X}}; \mathbf{w})$ and selecting the mean squared error (MSE) as the loss function, the gradient of the input samples is a multivariate polynomial with integer coefficients, i.e., $\boldsymbol{g}(\overline{\mathbf{X}}, \overline{\mathbf{Y}}; \mathbf{w}) \triangleq \nabla_{\mathbf{w}} \|\overline{\mathbf{Y}} - \boldsymbol{f}(\overline{\mathbf{X}}; \mathbf{w})\|_2^2 \in \mathbb{Z}^{d_w}$, where d_w represents the number of model parameters. In particular, to avoid wrap-around when computing gradient in the finite field $\mathbb{F}_p$, we assume the prime number p is sufficiently large without leading to overflow errors in the integer set $\mathbb{Z}_p$.

4 The Proposed DReS-FL Framework

DReS-FL consists of two main phases, as shown in Fig. 1. In the first phase, the private datasets are transformed from the real domain to the finite field, and data-owning clients secretly share datasets by Lagrange coding. Then, the server and the clients train a PINN iteratively via (1) local gradient computations and (2) gradient decoding and model updating.

4.1 Data Transformation and Secret Sharing

To guarantee information-theoretic privacy, each client has to mask the datasets in a finite field $\mathbb{F}_p$ using uniformly random matrices. Firstly, the local datasets $(\mathbf{X}_i, \mathbf{Y}_i)$ are converted from the real domain to the finite field $(\overline{\mathbf{X}}_i, \overline{\mathbf{Y}}_i)$. Considering an element-wise function $\phi(z) = z + c$ that transforms a real value to a non-negative number by adding a proper constant c[3], we define $\overline{\mathbf{X}} \triangleq Round(2^l \cdot \phi(\mathbf{X}))$, where the rounding operation is element-wise that quantizes each entry to its closest integer, and $l \in \mathbb{Z}$ controls the quantization loss.

After converting the private datasets to the finite field, data-owning clients adopt T-private Lagrange coding [31] to secretly share their local data samples across other clients, where T is the privacy parameter in our system ensuring that the encoded datasets do not leak any information about the original datasets even if T clients collude. First, each client $i \in [N]$ partitions its local dataset to K shards as $\overline{\mathbf{X}}^{(i)} = [\overline{\mathbf{X}}_1^{(i)T}, \ldots, \overline{\mathbf{X}}_K^{(i)T}]^T$ and $\overline{\mathbf{Y}}^{(i)} = [\overline{\mathbf{Y}}_1^{(i)T}, \ldots, \overline{\mathbf{Y}}_K^{(i)T}]^T$. Assuming that m_i is divisible by K, we have $\overline{\mathbf{X}}_k^{(i)T} \in \mathbb{F}_p^{\frac{m_i}{K} \times d_x}$ and $\overline{\mathbf{Y}}_k^{(i)T} \in \mathbb{F}_p^{\frac{m_i}{K} \times d_y}$ for $k \in [K]$. A large value of K helps to reduce the communication overhead in secret data sharing and computation costs in local gradient computations. Then, the clients add padding from T uniform random masks to the data samples for privacy protection. Each client $i \in [N]$ forms the following polynomials $\mathbf{u}_i(z)$ and $\mathbf{v}_i(z)$

[3]The constant scalar c could be the absolute value of the minimum entry in dataset $(\mathbf{X}, \mathbf{Y})$.

Algorithm 1 DReS-FL

Input: Local datasets $(\mathbf{X}_i, \mathbf{Y}_i)$ for $i \in [N]$, batch size b, initialized parameters $\mathbf{w}^{(0)} \in \mathbb{Z}_p^{d_w}$, distinct elements $\{\alpha_j\}_{j\in[N]}$ and $\{\beta_j\}_{j\in[K+T]}$, prime number p, training round τ, learning rate η.

Output: Model parameter $\mathbf{w}^{(\tau)}$.

Clients encode the local datasets according to (1) and (2) and deliver them to other clients.
for $t = 0, \ldots, \tau$ **do**
  Server sends the model parameters $\mathbf{w}^{(t)}$ to the clients.
  for $j = 1, \ldots, N$ **do**
    Client j performs gradient computation on mini-batches $(\widetilde{\mathbf{X}}_j^{(\mathcal{I}_t)}, \widetilde{\mathbf{Y}}_j^{(\mathcal{I}_t)})$.
    Upload stochastic gradient $\widetilde{\boldsymbol{g}}(\widetilde{\mathbf{X}}_j^{(\mathcal{I}_t)}, \widetilde{\mathbf{Y}}_j^{(\mathcal{I}_t)}; \mathbf{w}^{(t)})$ to the server.
  end for
  if Server receives at least $\deg(\boldsymbol{g})(K+T-1)+1$ uploads **then**
    Decode the gradients $\widetilde{\boldsymbol{g}}(\overline{\mathbf{X}}_j^{(\mathcal{I}_t)}, \overline{\mathbf{Y}}_j^{(\mathcal{I}_t)}; \mathbf{w}^{(t)})$ for $j \in [K]$ by polynomial interpolation.
    Convert gradients from the finite field to the integral domain $\boldsymbol{g}(\overline{\mathbf{X}}_j^{(\mathcal{I}_t)}, \overline{\mathbf{Y}}_j^{(\mathcal{I}_t)}; \mathbf{w}^{(t)})$ by (5).
    Update the global model by $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - Q(\frac{\eta}{bK}\sum_{j=1}^{K} \boldsymbol{g}(\overline{\mathbf{X}}_j^{(\mathcal{I}_t)}, \overline{\mathbf{Y}}_j^{(\mathcal{I}_t)}; \mathbf{w}^{(t)}))$ based on (6).
  end if
end for

of degree $K+T-1$:

$$\mathbf{u}_i(z) \triangleq \sum_{j\in[K]} \overline{\mathbf{X}}_j^{(i)} \cdot \prod_{k\in[K+T]\setminus\{j\}} \frac{z-\beta_k}{\beta_j-\beta_k} + \prod_{j=K+1}^{K+T} \mathbf{U}_j^{(i)} \cdot \prod_{k\in[K+T]\setminus\{j\}} \frac{z-\beta_k}{\beta_j-\beta_k}, \tag{1}$$

$$\mathbf{v}_i(z) \triangleq \sum_{j\in[K]} \overline{\mathbf{Y}}_j^{(i)} \cdot \prod_{k\in[K+T]\setminus\{j\}} \frac{z-\beta_k}{\beta_j-\beta_k} + \prod_{j=K+1}^{K+T} \mathbf{V}_j^{(i)} \cdot \prod_{k\in[K+T]\setminus\{j\}} \frac{z-\beta_k}{\beta_j-\beta_k}, \tag{2}$$

where $\{\mathbf{U}_j^{(i)}\}$'s and $\{\mathbf{V}_j^{(i)}\}$'s are random noise matrices uniformly sampled from $\mathbb{F}_p^{\frac{m}{K}\times d_x}$ and $\mathbb{F}_p^{\frac{m}{K}\times d_y}$, respectively. These matrices mask the local datasets and provide a privacy guarantee against up to T colluding workers. The clients and the server agree on $K+T$ distinct elements $\{\beta_1, \ldots, \beta_{K+T}\}$ from the finite field $\mathbb{F}_p$ in advance. Particularly, setting $z = \beta_k$ for $k \in [K]$, we obtain $\mathbf{u}_i(\beta_k) = \overline{\mathbf{X}}_k^{(i)}$ and $\mathbf{v}_i(\beta_k) = \overline{\mathbf{Y}}_k^{(i)}$. All the data-owning clients use the same N distinct elements $\{\alpha_1, \ldots, \alpha_N\}$ selected from $\mathbb{F}_p$ to encode the private dataset, where $\{\alpha_j\}_{j\in[N]} \cap \{\beta_j\}_{j\in[K+T]} = \varnothing$. Then, the encoded samples $(\widetilde{\mathbf{X}}_j^{(i)}, \widetilde{\mathbf{Y}}_j^{(i)}) \triangleq (\mathbf{u}_i(\alpha_j), \mathbf{v}_i(\alpha_j))$ are transmitted to client j from client i. All the received encoded datasets at client j are represented as $(\widetilde{\mathbf{X}}_j, \widetilde{\mathbf{Y}}_j)$, where $\widetilde{\mathbf{X}}_j \triangleq [\widetilde{\mathbf{X}}_j^{(1)T}, \ldots, \widetilde{\mathbf{X}}_j^{(N)T}]^T \in \mathbb{F}_p^{\widetilde{m}\times d_x}$ and $\widetilde{\mathbf{Y}}_j \triangleq [\widetilde{\mathbf{Y}}_j^{(1)T}, \ldots, \widetilde{\mathbf{Y}}_j^{(N)T}]^T \in \mathbb{F}_p^{\widetilde{m}\times d_y}$ for $j \in [N]$. Accordingly, the number of samples in the encoded dataset is $\widetilde{m} \triangleq \frac{1}{K}\sum_{i=1}^{n} m_i$.

4.2 Federated Training

Local Gradient Computation. The server randomly initializes a PINN at the beginning of the training process, and the model parameters are constrained to an integer set $\mathbb{Z}_p$ in the training process. In each communication round, the server sends the model parameters to the clients, and they compute the stochastic gradient over the mini-batches with size b. Particularly, we assume that all the clients use the same row selection matrix $\mathbf{C}^{(t)} \in \{0,1\}^{b\times\widetilde{m}}$ for data sampling in each round t[4], and the mini-batch at each client $j \in [N]$ is determined by $[\widetilde{\mathbf{X}}_j^{(\mathcal{I}_t)}, \widetilde{\mathbf{Y}}_j^{(\mathcal{I}_t)}] = \mathbf{C}^{(t)}[\widetilde{\mathbf{X}}_j, \widetilde{\mathbf{Y}}_j]$. Here, $\mathcal{I}_t = \{l_1^{(t)}, \ldots, l_b^{(t)}\} \subseteq [\widetilde{m}]$ is a randomly selected index set in the t-th round with $l_i \in [\widetilde{m}]$ for $i \in [b]$. The entries of $\mathbf{C}^{(t)}$ satisfy $\mathbf{C}_{i,l_i}^{(t)} = 1$ for $i \in [b]$, and other entries are set to zero. Each client j computes the stochastic gradient $\widetilde{\boldsymbol{g}}(\widetilde{\mathbf{X}}_j^{(\mathcal{I}_t)}, \widetilde{\mathbf{Y}}_j^{(\mathcal{I}_t)}; \mathbf{w}^{(t)})$ in the finite field, where

[4]This can be achieved by setting the same random seed across all the clients. The weighted sampling method has been adopted in this work, where the number of sampled data from $\widetilde{\mathbf{X}}_j^{(i)}$ is proportional to m_i for $i \in [N]$. Note that other sampling schemes can also be applied in DReS-FL.

$\widetilde{\boldsymbol{g}}(\widetilde{\mathbf{X}}_j^{(\mathcal{I}_t)}, \widetilde{\mathbf{Y}}_j^{(\mathcal{I}_t)}; \mathbf{w}^{(t)}) \equiv \boldsymbol{g}(\widetilde{\mathbf{X}}_j^{(\mathcal{I}_t)}, \widetilde{\mathbf{Y}}_j^{(\mathcal{I}_t)}; \mathbf{w}^{(t)}) \mod p$. Then, the local computations are uploaded to the server for gradient decoding, and each $\widetilde{\boldsymbol{g}}(\widetilde{\mathbf{X}}_j^{(\mathcal{I}_t)}, \widetilde{\mathbf{Y}}_j^{(\mathcal{I}_t)}; \mathbf{w}^{(t)})$ can be regarded as an evaluation of the polynomial $\widetilde{\boldsymbol{g}}(\mathbf{u}_{\mathcal{I}_t}(z), \mathbf{v}_{\mathcal{I}_t}(z); \mathbf{w}^{(t)})$ at the point α_j. The functions $\mathbf{u}_{\mathcal{I}_t}(z)$ and $\mathbf{v}_{\mathcal{I}_t}(z)$ are defined as follows:

$$\mathbf{u}_{\mathcal{I}_t}(z) \triangleq \sum_{j\in[K]} \overline{\mathbf{X}}_j^{(\mathcal{I}_t)} \cdot \prod_{k\in[K+T]\setminus\{j\}} \frac{z-\beta_k}{\beta_j-\beta_k} + \prod_{j=K+1}^{K+T} \mathbf{U}_j^{(\mathcal{I}_t)} \cdot \prod_{k\in[K+T]\setminus\{j\}} \frac{z-\beta_k}{\beta_j-\beta_k}, \tag{3}$$

$$\mathbf{v}_{\mathcal{I}_t}(z) \triangleq \sum_{j\in[K]} \overline{\mathbf{Y}}_j^{(\mathcal{I}_t)} \cdot \prod_{k\in[K+T]\setminus\{j\}} \frac{z-\beta_k}{\beta_j-\beta_k} + \prod_{j=K+1}^{K+T} \mathbf{V}_j^{(\mathcal{I}_t)} \cdot \prod_{k\in[K+T]\setminus\{j\}} \frac{z-\beta_k}{\beta_j-\beta_k}, \tag{4}$$

where

$$\overline{\mathbf{X}}_j^{(\mathcal{I}_t)} = \mathbf{C}^{(t)}[\overline{\mathbf{X}}_j^{(1)T}, \ldots, \overline{\mathbf{X}}_j^{(N)T}]^T \in \mathbb{F}_p^{b\times d_x}, \quad \overline{\mathbf{U}}_j^{(\mathcal{I}_t)} = \mathbf{C}^{(t)}[\mathbf{U}_j^{(1)T}, \ldots, \mathbf{U}_j^{(N)T}]^T \in \mathbb{F}_p^{b\times d_x},$$
$$\overline{\mathbf{Y}}_j^{(\mathcal{I}_t)} = \mathbf{C}^{(t)}[\overline{\mathbf{Y}}_j^{(1)T}, \ldots, \overline{\mathbf{Y}}_j^{(N)T}]^T \in \mathbb{F}_p^{b\times d_y}, \quad \overline{\mathbf{V}}_j^{(\mathcal{I}_t)} = \mathbf{C}^{(t)}[\mathbf{V}_j^{(1)T}, \ldots, \mathbf{V}_j^{(N)T}]^T \in \mathbb{F}_p^{b\times d_y}.$$

Gradient Decoding and Model Updating. After receiving the local gradients from any $\deg(\boldsymbol{g})(K+T-1)+1$ clients[5], the central server applies polynomial interpolation to perform gradient decoding. Particularly, the gradients $\widetilde{\boldsymbol{g}}(\widetilde{\mathbf{X}}_j^{(\mathcal{I}_t)}, \widetilde{\mathbf{Y}}_j^{(\mathcal{I}_t)}; \mathbf{w}^{(t)}), \forall j \in [N]$ amount to N evaluations of a composition of $(K+T-1)$-degree encoding polynomials with the gradient function $\boldsymbol{g}$. Therefore, the central server needs at least $\deg(\boldsymbol{g})(K+T-1)+1$ evaluations to interpolate the composed polynomial $\widetilde{\boldsymbol{g}}(\mathbf{u}_{\mathcal{I}_t}(z), \mathbf{v}_{\mathcal{I}_t}(z); \mathbf{w}^{(t)})$. Letting $z = \beta_j, \forall j \in [K]$, the server obtains K stochastic gradients $\widetilde{\boldsymbol{g}}(\overline{\mathbf{X}}_j^{(\mathcal{I}_t)}, \overline{\mathbf{Y}}_j^{(\mathcal{I}_t)}; \mathbf{w}^{(t)}) \equiv \widetilde{\boldsymbol{g}}(\mathbf{u}_{\mathcal{I}_t}(\beta_j), \mathbf{v}_{\mathcal{I}_t}(\beta_j); \mathbf{w}^{(t)})$ on the mini-batches $(\overline{\mathbf{X}}_j^{(\mathcal{I}_t)}, \overline{\mathbf{Y}}_j^{(\mathcal{I}_t)})$. Then, server converts the gradient from the finite field to the integer set $\mathbb{Z}_p$ by $\boldsymbol{g}(\overline{\mathbf{X}}_j^{(\mathcal{I}_t)}, \overline{\mathbf{Y}}_j^{(\mathcal{I}_t)}; \mathbf{w}^{(t)}) = \psi(\widetilde{\boldsymbol{g}}(\overline{\mathbf{X}}_j^{(\mathcal{I}_t)}, \overline{\mathbf{Y}}_j^{(\mathcal{I}_t)}; \mathbf{w}^{(t)}))$, where $\psi(z)$ is an element-wise function defined as follows:

$$\psi(z) = \begin{cases} z & \text{if} \quad 0 \le z < \frac{p-1}{2}, \\ z-p & \text{if} \quad \frac{p-1}{2} \le z < p. \end{cases} \tag{5}$$

As we assume that the prime number p is sufficiently large, the converted gradients do not have overflow errors. Thus, the central sever updates the global model by $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - Q(\frac{\eta}{bK}\sum_{j=1}^{K} \boldsymbol{g}(\overline{\mathbf{X}}_j^{(\mathcal{I}_t)}, \overline{\mathbf{Y}}_j^{(\mathcal{I}_t)}; \mathbf{w}^{(t)}))$, where η denotes the learning rate and bK represents the *global batch size*[6]. $Q(z)$ is a stochastic quantization function to ensure the model parameters are in the integer set $\mathbb{Z}_p$ after updating, which is defined as follows:

$$Q(z) = \begin{cases} \lfloor z \rfloor & \text{with probability } 1-(z-\lfloor z \rfloor) \\ \lfloor z \rfloor + 1 & \text{with probability } z - \lfloor z \rfloor. \end{cases} \tag{6}$$

Besides, the probability of rounding z to $\lfloor z \rfloor$ is proportional to the proximity of z to $\lfloor z \rfloor$ so that the stochastic rounding is unbiased. The overall procedure is summarized in Algorithm 1.

5 Theoretical Analysis

In this section, we characterize the theoretical performance of DReS-FL, in terms of (1) privacy and security guarantees, and (2) dropout-resiliency and convergence performance.

5.1 Privacy and Security Guarantees

Before training starts, client i receives an encoded version of the global dataset $(\widetilde{\mathbf{X}}_i, \widetilde{\mathbf{Y}}_i)$ from other clients. Lagrange coding in DReS-FL provides a strong privacy guarantee that the clients cannot infer

[5]Note that if the server cannot receive enough results due to the client dropouts, the training protocol continues to the next epoch without gradient decoding and model updating.

[6]The global batch size corresponds to the number of samples used for gradient computations in each round.

anything about the private datasets based on the received datasets even if up to T clients collude. The following theorem shows that our DReS-FL satisfies the first privacy requirement in Section 3, and the proof is available in Section IV of [31].

Theorem 1. *(T-private coding scheme [31]) Employing Lagrange coding with the privacy parameter T in DReS-FL, we have that for every subset of clients $\mathcal{T} \subseteq [N]$ of size at most T, the mutual information $I(\overline{\mathbf{X}}, \overline{\mathbf{Y}}; \{\widetilde{\mathbf{X}}_i, \widetilde{\mathbf{Y}}_i\}_{i\in\mathcal{T}}) = 0$.*

Besides, DReS-FL can provide a security guarantee such that the server learns no information from the local gradients beyond the global model $\mathbf{w}$. This property corresponds to the second privacy requirement in Section 3. The following theorem provides a rigorous statement and its proof is deferred to Appendix B.1.

Theorem 2. *(Security guarantee) For any $i \in [N]$ and index set $\mathcal{I}$, the conditional mutual information $I(\overline{\mathbf{X}}_i, \overline{\mathbf{Y}}_i; \widetilde{\boldsymbol{g}}(\widetilde{\mathbf{X}}_i^{(\mathcal{I})}, \widetilde{\mathbf{Y}}_i^{(\mathcal{I})}; \mathbf{w})|\mathbf{w})$ equals to zero.*

5.2 Dropout-resiliency and Convergence

In the FL setting, it is common for clients to drop out at any time during protocol execution, which leads to model divergence especially when the clients' datasets are highly heterogeneous. The following theorem shows that DReS-FL is resilient to a certain number of client dropouts.

Theorem 3. *(Dropout-resiliency) Consider N clients in the federated learning system that use a T-private coding scheme to secretly share the local samples $\overline{\mathbf{X}}^{(i)} = [\overline{\mathbf{X}}_1^{(i)T}, \ldots, \overline{\mathbf{X}}_K^{(i)T}]^T$ and $\overline{\mathbf{Y}}^{(i)} = [\overline{\mathbf{X}}_1^{(i)T}, \ldots, \overline{\mathbf{X}}_K^{(i)T}]^T$ for $i \in [N]$ for local gradient computations. DReS-FL guarantees that the server can decode the global gradient when there are no more than $D = N - \deg(\boldsymbol{g})(K+T-1)-1$ client dropouts.*

Proof. Theorem 1 of [31] shows that given a number of N computing nodes and a K-shard dataset, the LCC framework provides a T-private coding scheme for computing any polynomial $\boldsymbol{g}$, as long as $\deg(\boldsymbol{g})(K+T-1)+1 \leq N$. Thus, it can tolerate at most $N - \deg(\boldsymbol{g})(K+T-1)-1$ dropouts. □

Remark 1. There is a tradeoff among privacy guarantee (T), gradient computation cost ($1/K$), and dropout-resilience (D). Parameter T reflects the privacy threshold of Lagrange coding, and parameter K accounts for the computation load reduction. In particular, the local batch size of each client is $1/K$ of the global batch size. DReS-FL can achieve any T, K, and D as long as $D\deg(\boldsymbol{g})(K+T-1)+1 \leq N$. As T and K increase, DReS-FL can tolerate fewer client dropouts.

Remark 2. Our DReS-FL framework can be extended to more general cases in which clients can run s ($s \geq 1$) local SGD steps each round. Denote the computation results after s local SGD steps in round t as $\Delta\widetilde{\mathbf{w}}_i(s; \mathbf{w}^{(t)})$ for $i \in [N]$. Specifically, $\Delta\widetilde{\mathbf{w}}_i(s=1; \mathbf{w}^{(t)}) = \widetilde{\boldsymbol{g}}(\overline{\mathbf{X}}_i^{(\mathcal{I}_t)}, \overline{\mathbf{Y}}_i^{(\mathcal{I}_t)}; \mathbf{w}^{(t)})$ and $\Delta\widetilde{\mathbf{w}}_i(s=2; \mathbf{w}^{(t)}) = \widetilde{\boldsymbol{g}}(\overline{\mathbf{X}}_i^{(\mathcal{I}_{t+1})}, \overline{\mathbf{Y}}_i^{(\mathcal{I}_{t+1})}; \mathbf{w}^{(t)} - \frac{\eta}{bK}\widetilde{\boldsymbol{g}}(\overline{\mathbf{X}}_i^{(\mathcal{I}_t)}, \overline{\mathbf{Y}}_i^{(\mathcal{I}_t)}; \mathbf{w}^{(t)}))$. By carefully selecting the learning rate η such that $\frac{\eta}{bK} \in \mathbb{F}_p$, the function $\Delta\widetilde{\mathbf{w}}_i(s)$ is still a polynomial in the finite field $\mathbb{F}_p$. Therefore, the central server can recover the desired model update by polynomial interpolation at the cost of low dropout-resiliency caused by the high degree of $\Delta\widetilde{\mathbf{w}}_i(s; \mathbf{w}^{(t)})$.

Next, we characterize the convergence performance of PINNs, which relies on the fact that the global gradients in the training process are unbiased. Denote the true loss function as $\ell(\mathbf{w}) \triangleq \mathbb{E}_{\overline{\mathbf{X}},\overline{\mathbf{Y}}}\|\overline{\mathbf{Y}} - \boldsymbol{f}(\overline{\mathbf{X}}; \mathbf{w})\|_2^2$ and the corresponding gradient as $\boldsymbol{g}_e(\mathbf{w}) \triangleq \mathbb{E}_{\overline{\mathbf{X}},\overline{\mathbf{Y}}}[\frac{1}{m}\boldsymbol{g}(\overline{\mathbf{X}}, \overline{\mathbf{Y}}; \mathbf{w})]$. To prove that DReS-FL guarantees convergence to the optimal model parameters, we first present the following amputations to facilitate the convergence analysis.

Assumption 1. *(L-smoothness) There exists a constant $L > 0$ such that for all $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{Z}_p^{d_w}$, we have $\|\boldsymbol{g}_e(\mathbf{w}_1) - \boldsymbol{g}_e(\mathbf{w}_2)\|_2 \leq L\|\mathbf{w}_1 - \mathbf{w}_2\|_2$.*

Assumption 2. *(Unbiased and variance-bounded stochastic gradient) There exists a constant $\sigma > 0$ such that any stochastic gradient $\boldsymbol{g}(\widetilde{\mathbf{X}}_j^{(\mathcal{I}_t)}, \widetilde{\mathbf{Y}}_j^{(\mathcal{I}_t)}; \mathbf{w}^{(t)})$ satisfies $\mathbb{E}\left[\boldsymbol{g}(\widetilde{\mathbf{X}}_j^{(\mathcal{I}_t)}, \widetilde{\mathbf{Y}}_j^{(\mathcal{I}_t)}; \mathbf{w}^{(t)})\right] = \boldsymbol{g}_e(\mathbf{w}^{(t)})$ and $\mathbb{E}\left[\|\boldsymbol{g}(\widetilde{\mathbf{X}}_j^{(\mathcal{I}_t)}, \widetilde{\mathbf{Y}}_j^{(\mathcal{I}_t)}; \mathbf{w}^{(t)}) - \boldsymbol{g}_e(\mathbf{w}^{(t)})\|^2\right] \leq \sigma^2$.*

Table 1: Test accuracy $(\%)$ of different methods. Each experiment is repeated five times.

Dataset	MNIST	Fashin-MNIST	EMNIST	CIFAR-10	CIFAR-100	SVHN
FedAvg	96.17 ± 0.05	81.20 ± 0.07	71.50 ± 0.28	89.54 ± 0.09	67.71 ± 0.26	83.82 ± 0.20
FedAvg-IS	97.06 ± 0.10	85.94 ± 0.16	77.09 ± 0.34	89.83 ± 0.07	68.92 ± 0.14	85.27 ± 0.09
SCAFFOLD	71.89 ± 3.92	55.22 ± 1.83	55.15 ± 5.95	54.17 ± 9.13	29.97 ± 1.73	51.27 ± 3.43
DReS-FL (Ours)	$\mathbf{97.38 \pm 0.08}$	$\mathbf{86.60 \pm 0.32}$	$\mathbf{78.04 \pm 0.29}$	$\mathbf{90.31 \pm 0.19}$	$\mathbf{69.15 \pm 0.27}$	$\mathbf{86.04 \pm 0.15}$
Centralized	97.99 ± 0.04	89.02 ± 0.11	82.45 ± 0.23	90.37 ± 0.12	71.12 ± 0.09	86.18 ± 0.03

Assumption 3. *(Unbiased and variance-bounded rounding operation) There exists a constant $\gamma > 0$ such that for any $z \in \mathbb{R}$, the stochastic quantization operation $Q(\cdot)$ satisfies $\mathbb{E}\left[Q(z)\right] = z$ and $\mathbb{E}\left[\|Q(z) - z\|^2\right] \leq \gamma^2 z^2$.*

With the above preparations, we have the following theorem which ensures the convergence. The proof is deferred to Appendix B.2.

Theorem 4. *(Convergence) With Assumption 1-3, selecting the learning rate as $\eta = \mathcal{O}\left(1/\sqrt{\tau'}\right)$ such that $\Psi = 1 - \eta L/2 - \eta\gamma^2 L/2 > 0$, after τ' times of model updates, we have:*

$$\frac{1}{\tau'}\sum_{t=1}^{\tau'} \|\boldsymbol{g}_e(\mathbf{w}^{(t)})\|^2 \leq \frac{\ell(\mathbf{w}^{(0)}) - \ell(\mathbf{w}^*)}{\eta\tau'\Psi} + \frac{\eta^2 L\sigma^2}{2bK\Psi}(\gamma^2 + 1), \tag{7}$$

where $\mathbf{w}^*$ is the first-order optimal solution.

6 Experiments

We demonstrate the performance of DReS-FL under the setting where local datasets are heterogeneous and clients may drop out of the training process arbitrarily.

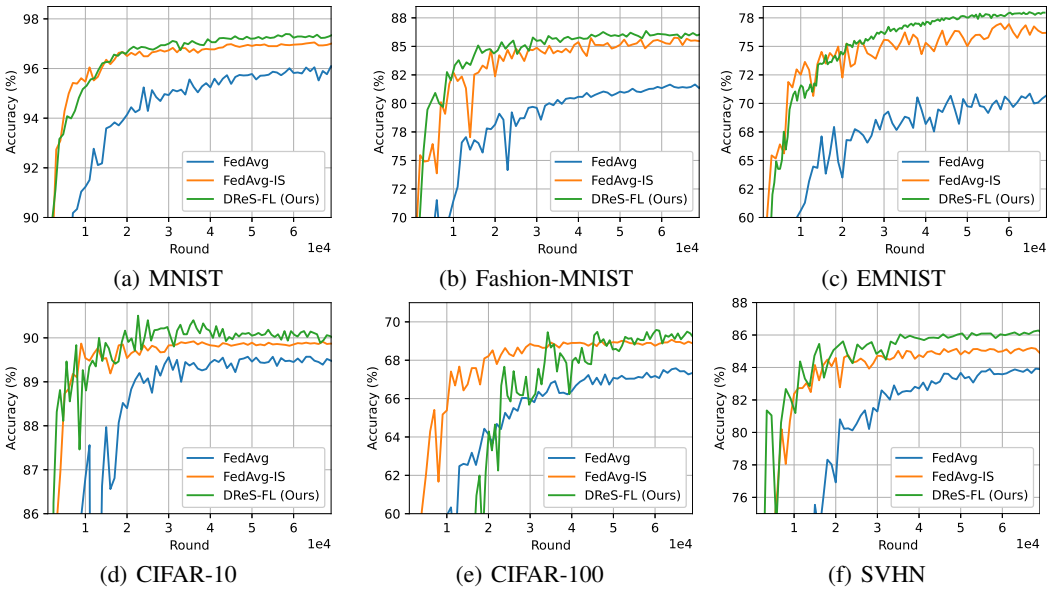
6.1 Experimental Setup

Dataset. We evaluate our proposed algorithm on several benchmark datasets: MNIST [45], Fashion-MNIST [46], EMNIST (Balanced) [47], CIFAR-10 [48], CIFAR-100 [48], and SVHN [49]. To simulate the non-IID data distribution, we assume there are $N = 20$ clients in the learning system and adopt the skewed label partition [50] to shuffle the datasets. Specifically, we sort a dataset by the labels, divide it into N shards, and assign one shard to each client. To simulate the client dropouts in the training process, we consider an extreme scenario, where the dropout rate of each client is set to 0.99 with a probability of 0.5 or is uniformly sampled from $[0, 0.1]$ otherwise.

Model structures. We adopt a multi-layer perception (MLP) with two hidden layers for the image classification tasks on MNIST, Fashion-MNIST, and EMNIST datasets. Besides, we utilize an ImageNet pretrained VGG19 model [51, 52] for CIFAR-10, CIFAR-100, and SVHN datasets. Specifically, the parameters in convolutional layers of VGG19 are fixed, and we utilize an MLP with two hidden layers as a classifier. The baseline methods train the neural networks on the real field and select the rectified linear unit (ReLU) function as the activation function. In each communication round, clients perform one SGD step for the local model update.

DReS-FL. Our method adopts the same size PINNs to replace MLPs in the federated training, and the degree of gradient is $\deg(\boldsymbol{g}) = 8$. Particularly, the extracted features from the last convolutional layer of VGG19 are secretly shared with other clients. We set the parameters $K = 1$ and $T = 1$ in the Lagrange coding, and the minimum number of clients needed to decode the global gradient is 9.

Baselines. We select algorithm-based methods as baselines, including FedAvg [1], FedAvg with importance sampling (FedAvg-IS) [53, 54], and SCAFFOLD [5]. These methods can be easily combined with secure aggregation methods [9, 13] to satisfy the privacy requirements identified in Section 3. Specifically, we assume that the FedAvg-IS method knows the dropout distribution, and the local updates are weighted by the dropout probabilities to mitigate bias. Besides, we adopt the centralized training scheme as a performance upper bound in the comparison. More details of the experimental settings are deferred to Appendix A.



(a) MNIST (b) Fashion-MNIST (c) EMNIST

(d) CIFAR-10 (e) CIFAR-100 (f) SVHN

Figure 2: Test accuracies on different datasets.

6.2 Performance Evaluation

We compare the performance of our DReS-FL method with baselines. The experimental results are shown in Table 1 and Fig. 2. We observe that FedAvg is influenced by the dropout problem due to the biased gradients. The FedAvg-IS method improves the test accuracy compared with FedAvg, but there is still a noticeable performance gap with the centralized training scheme. It shows that using the knowledge of dropout distribution can partially compensate for the biases in the aggregated models, but the local data distributions are still heterogeneous and degrade the performance. Besides, SCAFFOLD has a low accuracy on all the settings. As the frequency of updating local control variates is low, the estimation of the update direction is highly inaccurate such that the model does not converge as shown in Fig. 3. These results are consistent with the findings in [19]. Our DReS-FL method is superior to all the baseline methods as the server can obtain global gradients after polynomial interpolation. In addition, DReS-FL achieves comparable performance to the centralized training scheme on some datasets, which demonstrates the effectiveness of our proposed framework in solving the non-IID and dropout problems.

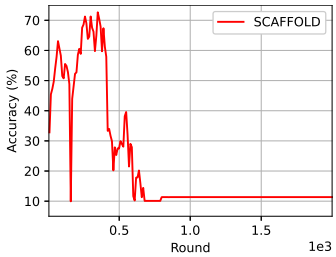


Figure 3: Test accuracy of SCAFFOLD on MNIST.

7 Conclusions and Discussions

This paper proposed a Dropout-Resilient Secure Federated Learning (DReS-FL) framework via Lagrange coded computing (LCC) to simultaneously solve the data heterogeneity and dropout problems of FL, while providing privacy guarantees for the local datasets. The polynomial integer neural networks (PINNs) have been constructed to ensure that the server can correctly decode the global gradient without privacy leakage. Extensive experimental results validated the effectiveness of the proposed method. Potential limitations of our method include that the degree of the gradient in a PINN increases exponentially with the number of layers, which hinders training a deep model for complex tasks. Besides, performing multiple local SGD steps largely increases the finite field size as the range of results grows exponentially with the number of multiplications, and thus it will lead to substantial communication overhead in model transmission. Despite some limitations, we believe DReS-FL is a promising framework for many practical FL application scenarios given its effectiveness in resolving both the non-IID and client dropout problems, while with strong privacy guarantees.

References

[1] McMahan, B., E. Moore, D. Ramage, et al. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[2] Shokri, R., M. Stronati, C. Song, et al. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[3] Nasr, M., R. Shokri, A. Houmansadr. Comprehensive privacy analysis of deep learning. In *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)*, pages 1–15. 2018.

[4] Geiping, J., H. Bauermeister, H. Dröge, et al. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33:16937–16947, 2020.

[5] Karimireddy, S. P., S. Kale, M. Mohri, et al. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

[6] Kairouz, P., H. B. McMahan, B. Avent, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[7] Hsieh, K., A. Phanishayee, O. Mutlu, et al. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pages 4387–4398. PMLR, 2020.

[8] Luo, M., F. Chen, D. Hu, et al. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. *Advances in Neural Information Processing Systems*, 34, 2021.

[9] Bonawitz, K. and Ivanov, Vladimir and Kreuter, Ben and Marcedone, Antonio and McMahan, H Brendan and Patel, Sarvar and Ramage, Daniel and Segal, Aaron and Seth, Karn. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191. 2017.

[10] Bonawitz, K., V. Ivanov, B. Kreuter, et al. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191. 2017.

[11] So, J., B. Güler, A. S. Avestimehr. Turbo-aggregate: Breaking the quadratic aggregation barrier in secure federated learning. *IEEE Journal on Selected Areas in Information Theory*, 2(1):479–489, 2021.

[12] Kadhe, S., N. Rajaraman, O. O. Koyluoglu, et al. Fastsecagg: Scalable secure aggregation for privacy-preserving federated learning. *arXiv preprint arXiv:2009.11248*, 2020.

[13] Yang, C.-S., J. So, C. He, et al. Lightsecagg: Rethinking secure aggregation in federated learning. *arXiv preprint arXiv:2109.14236*, 2021.

[14] Jahani-Nezhad, T., M. A. Maddah-Ali, S. Li, et al. Swiftagg+: Achieving asymptotically optimal communication load in secure aggregation for federated learning. *arXiv preprint arXiv:2203.13060*, 2022.

[15] Sahu, A. K., T. Li, M. Sanjabi, et al. On the convergence of federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 3:3, 2018.

[16] Li, Q., B. He, D. Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10713–10722. 2021.

[17] Acar, D. A. E., Y. Zhao, R. M. Navarro, et al. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.

[18] Hsu, T.-M. H., H. Qi, M. Brown. Federated visual classification with real-world data distribution. In *European Conference on Computer Vision*, pages 76–92. Springer, 2020.

[19] Li, Q., Y. Diao, Q. Chen, et al. Federated learning on non-iid data silos: An experimental study. *arXiv preprint arXiv:2102.02079*, 2021.

[20] Zhao, Y., M. Li, L. Lai, et al. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

[21] Yoshida, N., T. Nishio, M. Morikura, et al. Hybrid-fl for wireless networks: Cooperative learning mechanism using non-iid data. In *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, pages 1–7. IEEE, 2020.

[22] Yoon, T., S. Shin, S. J. Hwang, et al. Fedmix: Approximation of mixup under mean augmented federated learning. In *International Conference on Learning Representations*. 2020.

[23] Sun, Y., J. Shao, S. Li, et al. Stochastic coded federated learning with convergence and privacy guarantees. *arXiv preprint arXiv:2201.10092*, 2022.

[24] Jeong, E., S. Oh, H. Kim, et al. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. *arXiv preprint arXiv:1811.11479*, 2018.

[25] Hao, W., M. El-Khamy, J. Lee, et al. Towards fair federated learning with zero-shot data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3310–3319. 2021.

[26] Zhang, L., B. Shen, A. Barnawi, et al. Feddpgan: federated differentially private generative adversarial networks framework for the detection of covid-19 pneumonia. *Information Systems Frontiers*, 23(6):1403–1415, 2021.

[27] Nguyen, D. C., M. Ding, P. N. Pathirana, et al. Federated learning for covid-19 detection with generative adversarial networks in edge cloud computing. *IEEE Internet of Things Journal*, 2021.

[28] Jeong, E., S. Oh, H. Kim, et al. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. *arXiv preprint arXiv:1811.11479*, 2018.

[29] Zhu, Z., J. Hong, J. Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *International Conference on Machine Learning*, pages 12878–12889. PMLR, 2021.

[30] Li, Z., J. Shao, Y. Mao, et al. Federated learning with GAN-based data synthesis for non-IID clients, 2022.

[31] Yu, Q., S. Li, N. Raviv, et al. Lagrange coded computing: Optimal design for resiliency, security, and privacy. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1215–1225. PMLR, 2019.

[32] Shamir, A. How to share a secret. *Communications of the ACM*, 22(11):612–613, 1979.

[33] Gentry, C. Fully homomorphic encryption using ideal lattices. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 169–178. 2009.

[34] Gilad-Bachrach, R., N. Dowlin, K. Laine, et al. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *International conference on machine learning*, pages 201–210. PMLR, 2016.

[35] Hesamifard, E., H. Takabi, M. Ghasemi. Cryptodl: towards deep learning over encrypted data. In *Annual Computer Security Applications Conference (ACSAC 2016), Los Angeles, California, USA*, vol. 11. 2016.

[36] Graepel, T., K. Lauter, M. Naehrig. Ml confidential: Machine learning on encrypted data. In *International Conference on Information Security and Cryptology*, pages 1–21. Springer, 2012.

[37] Yuan, J., S. Yu. Privacy preserving back-propagation neural network learning made practical with cloud computing. *IEEE Transactions on Parallel and Distributed Systems*, 25(1):212–221, 2013.

[38] Han, K., S. Hong, J. H. Cheon, et al. Logistic regression on homomorphic encrypted data at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pages 9466–9471. 2019.

[39] Wang, Q., M. Du, X. Chen, et al. Privacy-preserving collaborative model learning: The case of word vector training. *IEEE Transactions on Knowledge and Data Engineering*, 30(12):2381–2393, 2018.

[40] Nikolaenko, V., U. Weinsberg, S. Ioannidis, et al. Privacy-preserving ridge regression on hundreds of millions of records. In *2013 IEEE symposium on security and privacy*, pages 334–348. IEEE, 2013.

[41] Gascón, A., P. Schoppmann, B. Balle, et al. Privacy-preserving distributed linear regression on high-dimensional data. *Proc. Priv. Enhancing Technol.*, 2017(4):345–364, 2017.

[42] Mohassel, P., Y. Zhang. Secureml: A system for scalable privacy-preserving machine learning. In *2017 IEEE symposium on security and privacy (SP)*, pages 19–38. IEEE, 2017.

[43] So, J., B. Güler, A. S. Avestimehr. Codedprivateml: A fast and privacy-preserving framework for distributed machine learning. *IEEE Journal on Selected Areas in Information Theory*, 2(1):441–451, 2021.

[44] So, J., B. Guler, S. Avestimehr. A scalable approach for privacy-preserving collaborative machine learning. *Advances in Neural Information Processing Systems*, 33:8054–8066, 2020.

[45] LeCun, Y., L. Bottou, Y. Bengio, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[46] Xiao, H., K. Rasul, R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[47] Cohen, G., S. Afshar, J. Tapson, et al. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

[48] Krizhevsky, A., G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[49] Netzer, Y., T. Wang, A. Coates, et al. Reading digits in natural images with unsupervised feature learning. 2011.

[50] Hsieh, K., A. Phanishayee, O. Mutlu, et al. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pages 4387–4398. PMLR, 2020.

[51] Simonyan, K., A. Zisserman. Very deep convolutional networks for large-scale image recognition. 2015.

[52] Krizhevsky, A., I. Sutskever, G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[53] Ren, J., Y. He, D. Wen, et al. Scheduling for cellular federated edge learning with importance and channel awareness. *IEEE Transactions on Wireless Communications*, 19(11):7690–7703, 2020.

[54] Kairouz, P., H. B. McMahan, B. Avent, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 7
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] The proofs are provided in Section 5 and Appendix B.
3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] The code is proprietary, but the public datasets used in the experiments and the instructions needed for reproduction are presented in Section 6 and Appendix A.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Please refer to Section 6 and Appendix A.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 6.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix A.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [Yes] See Appendix A.

(c) Did you include any new assets either in the supplemental material or as a URL? [No] The code is proprietary and the datasets used in our work are public datasets.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We choose the commonly-used datasets in the experiments.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]