

# Immersion and Invariance-based Coding for Privacy-Preserving Federated Learning<sup>★</sup>

Haleh Hayati<sup>a</sup> Carlos Murguia<sup>a,b</sup> Nathan van de Wouw<sup>a</sup>

<sup>a</sup>*Dynamics & Control Group, Department of Mechanical Engineering, Eindhoven University of Technology, The Netherlands.*

<sup>b</sup>*Engineering Systems and Design, Singapore University of Technology and Design, Singapore.*

---

## Abstract

Federated learning (FL) has emerged to preserve privacy in collaborative distributed learning. In FL, clients conduct AI model training directly on their devices rather than sharing their data with a centralized server, which could potentially pose privacy risks. However, it has been shown that despite FL's partial preservation of local data privacy, information about clients' data can still be inferred from shared model updates during the training process. In recent years, several privacy-preserving approaches have been developed to mitigate this privacy leakage in FL. However, they often provide privacy at the cost of model performance or system efficiency. Balancing these trade-offs poses a significant challenge in implementing FL schemes. In this manuscript, we introduce a privacy-preserving FL framework built on the synergy of differential privacy and system immersion and invariance tools from control theory. The core idea is to treat optimization algorithms used in the standard FL schemes (gradient-based algorithms) as a dynamical system that we seek to immerse into a higher-dimensional system (referred here to as the target optimization algorithm). The dynamics of the target optimization algorithm is designed such that, firstly, the model parameters of the original algorithm are immersed/embedded in its parameters, secondly, it works on distorted parameters, and, thirdly, converges to an encoded version of the true model parameters of the original algorithm. The encoded model parameters can be decoded at the server to extract the original model parameters. We demonstrate that the proposed privacy-preserving scheme can be tailored to offer any desired level of differential privacy for local and global model parameters while maintaining the same accuracy and convergence rate as standard FL algorithms.

*Key words:* Privacy-preservation, Federated Learning, Immersion and Invariance, Differential Privacy.

---

## 1 Introduction

Machine learning (ML) has been successfully applied in various applications for multiple fields and industries. In traditional machine learning, the server conducting the learning algorithm typically holds the training data centrally. However, when multiple participants are involved, sharing local data with the server poses a significant privacy risk, given the potential disclosure of private information. To address this issue, Federated learning (FL) [1, 2] has been introduced as a decentralized learning framework, enabling collaboration among numerous participants while preserving data privacy. Its fundamental concept involves training ML models on separate databases distributed across several devices or entities. FL schemes train local models on local clients'

databases, with clients subsequently sharing their parameters (e.g., model weights or gradients) with a central server to aggregate a global model. FL is suitable for sensitive data sharing, e.g., in the scope of healthcare and the Internet of Things (IoT), because clients do not need to directly share their training data [3]. However, despite FL's efforts to preserve clients' raw data privacy by avoiding direct data exchange, research has proven that private information can still be inferred from model parameters throughout the training process. It has been shown that local models can be traced back to their sources [4]. Additionally, private information can be extracted from multiple aggregated global models at the central server [5, 6]. Common attacks to FL are model inversion attacks and gradient inference attacks [7, 8]. In recent years, various approaches have been developed to achieve Privacy-Preserving FL (PPFL) [9]. Most of them rely on perturbation-based techniques such as Differential Privacy [10–14], and cryptography-based techniques such as Secure Multi-Party Computation [15–18], and Homomorphic Encryption [19–22]. Differential Privacy (DP) offers strong probabilistic privacy guarantees with minimal system overhead and algorithmic

---

<sup>★</sup> This work is supported by European Union's Horizon Europe programme under grant agreement No 101069748 – SELFY project.

*Email addresses:* h.hayati@tue.nl (Haleh Hayati),  
c.g.murguia@tue.nl (Carlos Murguia),  
n.v.d.wouw@tue.nl (Nathan van de Wouw).

simplicity. However, it introduces a trade-off between privacy and FL performance, as the added noise can significantly degrade model accuracy and slow convergence. Secure Multi-Party Computation (MPC) allows distributed clients to jointly compute a function without revealing their individual inputs, making it suitable for privacy-preserving model aggregation in FL. While MPC avoids the accuracy loss in DP, it demands additional communication between the server and clients, leading to increased communication costs and overhead. Homomorphic Encryption (HE), another cryptographic method, enables computations on encrypted data. With this approach, clients send their encrypted models to the server to aggregate them without decryption. While HE-based FL preserves model accuracy and eliminates complex client interactions, it is computationally intensive, leading to significant computational overhead. Both cryptographic approaches face challenges related to high communication costs and computational demands. Additionally, while these methods ensure the private aggregation of local models, they are vulnerable to inference attacks over the aggregated models. To address the limitations of cryptographic techniques and DP, recent efforts have focused on hybrid methods combining cryptographic tools with DP schemes to hold acceptable trade-offs between data privacy and FL performance [23–26]. Although existing techniques improve privacy of FL, they often do it at the cost of model performance and system efficiency. Balancing these trade-offs is challenging when implementing private FL systems. Therefore, novel PPFL schemes must be designed to provide strict privacy guarantees with a fair computational cost without compromising the model performance excessively. The aim of this work is to design coding mechanisms that protect the private information of local and global models while implementing FL algorithms. We propose a PPFL framework built on the synergy of random coding and *system immersion* tools [27] from control theory. The core idea involves treating the Gradient descent optimization algorithms, e.g., SGD, Adam, Momentum, etc., commonly used in standard FL, as a dynamical system that we seek to *immerse* into a higher-dimensional algorithm (the so-called target optimization algorithm). Essentially, this means that model parameters in the target optimizer must embed all model parameters of the original optimizer (up to a random invertible transformation). The target optimization algorithm must be designed so that: 1) model parameters of the original algorithm are immersed/embedded in its parameters, and 2) it operates on randomly encoded higher-dimensional parameters to produce randomly encoded optimal model parameters. We formulate a coding mechanism at the server side as a random change of coordinates that maps original model parameters to a higher-dimensional parameter space. Such coding enforces that the target optimization algorithm at the clients’ side converges to an encoded higher-dimensional version of the optimal model parameters of the standard algorithm. The encoded local model parameters are aggregated by a third

party (the so-called aggregator). Another coding is formulated for the aggregator to encode the aggregated model to avoid the server’s access to intermediate global models (since the server has access to the first encoding and decoding maps). The aggregator transmits the aggregated encoded global model to the server. The aggregated model is decoded at the server side using the left inverse of the server encoding map (the decoded model at the server side is still encoded by the aggregator coding). Since the aggregator only has access to the encoded local models it does not need to be trusted. In addition, the server only has access to the encoded aggregated model by the aggregator and does not need to be trusted.

The general idea of system immersion-based coding is introduced in [28] as a *homomorphic encryption scheme that operates over the reals* aiming to preserve privacy of centralized dynamical algorithms. Compared to standard HE, this approach offers much more freedom to redesign algorithms to work on the encoded/encrypted data. The immersion-based coding scheme provides the same utility as the original algorithm (i.e., when no coding is employed), is computationally efficient, can be applied to large-scale algorithms, and gives arbitrarily strong privacy probabilistic guarantees (in terms of differential privacy) without degrading the algorithm performance. We have started exploring these ideas in [29,30] to protect privacy of local models in the aggregation steps of FL algorithms. However, it has been shown that the server can recover local models by accessing multiple intermediate global models [6]. Therefore, it is necessary to provide privacy against inference attacks over both local and intermediate global models. To address this, we propose employing the immersion-based coding idea for privacy of intermediate local and global models in FL by encoding model parameters at both the server and the aggregator at every iteration. This prevents all internal parties involved, server, clients, aggregator, and external parties such as model consumers and eavesdroppers, from accessing actual models. We generalize this coding approach for various machine learning and deep learning models to cover gradient descent optimizers for different applications. Simulation experiments and rigorous mathematical proofs indicate that our framework maintains the same accuracy and convergence rate as standard FL, reveals no information about the clients’ data, and is computationally efficient.

The main contributions of the paper are as follows:

- Using random coding and *system immersion* tools, we develop a prescriptive synthesis framework for the design of a privacy-preserving FL algorithm that guarantees data privacy against the inversion of local and global models in the aggregation and broadcasting steps of FL. This extends beyond our previous work in [29] and other cryptographic tools, which only consider privacy in the aggregation of local models.
- We demonstrate that the proposed scheme provides any desired level of differential privacy guarantee for local and global models without compromising the ac-

Table 1

Theoretical comparison of existing privacy-preserving FL frameworks.

Approach	No accuracy loss	No communication costs	No computation overhead	Aggregated model privacy
DP [10–13]	✗	✓	✓	✓
MPC [15–18]	✓	✗	✗	✗
HE [19–21]	✓	✓	✗	✗
Ours	✓	✓	✓	✓

Table 2

Description of Main Notation.

Symbol	Description
$K$	Number of local iterations for clients
$T$	Number of global iterations of FL
$N_c$	Number of clients
$\mathcal{D}, \mathcal{D}'$	Adjacent databases
$\mathcal{C}_i$	$i^{\text{th}}$ client
$\mathcal{D}_i$	The database held by $\mathcal{C}_i$
$w_{i,k}$	$\mathcal{C}_i$ 's local parameters at $k^{\text{th}}$ local iteration
$w_{i,K}$	Local uploading parameters of client $\mathcal{C}_i$
$w^t$	Global aggregated parameters from all local parameters at the $t^{\text{th}}$ global iteration
$\tilde{w}^t$	Global aggregated parameters encoded by the server at $t^{\text{th}}$ global iteration
$\bar{w}^t$	Global aggregated parameters encoded by the aggregator at $t^{\text{th}}$ global iteration
$w'^t$	Global aggregated parameters encoded by the aggregator and server at $t^{\text{th}}$ global iteration
$\pi_1(\cdot)$	Server encoding map
$\pi_1^L(\cdot)$	Server decoding map (left inverse of $\pi_1(\cdot)$ )
$\pi_2(\cdot)$	Aggregator encoding map
$\pi_2^R(\cdot)$	Clients decoding map (right inverse of $\pi_2(\cdot)$ )
$n$	Number of model parameters
$\tilde{n}$	Number of parameters encoded by server
$\bar{n}$	Number of parameters encoded by aggregator
$n'$	Number of parameters after encoding by aggregator and server

curacy and convergence rate of the federated learning algorithm with a fair computation cost, in contrast to other related approaches that often impact model performance or system efficiency. This guarantee is not provided in our preliminary work in [29].

- We validate the effectiveness of the scheme through extensive computer simulations and illustrate that it is possible to employ the proposed scheme to train a variety of ML models without affecting their accuracy and convergence rate through experimental evaluation. Therefore, our analytical results are helpful for the design of privacy-preserving FL architectures with a variety of ML networks and settings.

Based on the previous discussion, summarized in Table 1, our proposed PPFL approach is advantageous in terms of model performance, system efficiency, and privacy of intermediate aggregated models.

The notation of the paper is introduced in Table 2.

## 2 Problem Formulation

### 2.1 Standard Federated Learning

Our scheme’s architecture is developed by expanding upon the standard FL algorithm. Within the standard

FL framework, the objective is to train a global AI model collaboratively across multiple dispersed devices, referred to as clients, and a central server while avoiding the direct exchange of local data held by the clients. Instead, clients transmit their local model parameters to the server, obtained through training on their devices with local data. The server then aggregates these local models’ parameters to construct a comprehensive global model, which is subsequently sent back to the clients. These updated global parameters serve as the initial conditions for clients to refine their local models. This iterative process continues until convergence [31]. We consider a standard FL system consisting of one server and  $N_c$  clients. The local database owned by the  $i^{\text{th}}$  client,  $\mathcal{C}_i$ ,  $i \in \{1, 2, \dots, N_c\}$ , is denoted by  $\mathcal{D}_i$ . At each iteration  $t \in \mathbb{N}$ , the server broadcasts the latest global model,  $w^t \in \mathbb{R}^n$  (a vector of parameters), to all clients (beginning from a random initial model  $w^0$ ). These iteration times  $t$  are termed as global iterations. Next, clients use the latest update on  $w^t$  and local data  $\mathcal{D}_i$  to minimize a specified loss function  $l(w_i^t, \mathcal{D}_i)$  at their devices to identify local AI models,  $w_i^t \in \mathbb{R}^n$ . This can be written as follows:

$$w_i^{t+1} = \arg \min_{w_i^t} l(w_i^t, \mathcal{D}_i). \quad (1)$$

The clients transmit their local optimal  $w_i^{t+1}$  back to the server, which then proceeds to update the global model as follows:

$$w^{t+1} = \sum_{i=1}^{N_c} \frac{|\mathcal{D}_i|}{|\mathcal{D}|} w_i^{t+1}, \quad (2)$$

where  $|\mathcal{D}_i|$  is the size of the  $i^{\text{th}}$  database,  $|\mathcal{D}| := \sum_i |\mathcal{D}_i|$ , and  $w^t$  is the global aggregated model. The procedure iterates until reaching convergence towards the global optimum (see [1]):

$$w^* = \arg \min_w \sum_{i=1}^{N_c} \frac{|\mathcal{D}_i|}{|\mathcal{D}|} l(w, \mathcal{D}_i). \quad (3)$$

In general, standard FL clients use gradient descent optimization algorithms [32], e.g., Stochastic Gradient Descent (SGD), Adam, etc., as the optimization tool to minimize their local loss function (1). Each client calculates the stochastic gradient of the local model using their local data  $\mathcal{D}_i$  and updates its local model following  $K$  iterations of the optimizers that can generally be

modeled as follows:

$$\text{Optimizer} \begin{cases} w_{i,0} = w^t, \\ w_{i,k+1} = f(w_{i,k}, \mathcal{D}_i) := w_{i,k} - g(w_{i,k}, \mathcal{D}_i), \\ k = 0, 1, \dots, K-1, \\ w_i^{t+1} = w_{i,K}, \end{cases} \quad (4)$$

where  $w_{i,k} \in \mathbb{R}^n$  denotes the local model parameters in the  $k^{\text{th}}$  local iteration of the optimization algorithm at the  $i^{\text{th}}$  client, and  $g(w_{i,k}, \mathcal{D}_i)$  is a gradient-based function, reflecting the step in parameters, that can be defined for each gradient descent optimizer.

Stochastic Gradient Descent (SGD) is a common gradient descent optimization technique in FL [32]. For SGD, function  $g(\cdot, \cdot)$  can be written as  $g(w_{i,k}, \mathcal{D}_i) := \eta \nabla l(w_{i,k}, \mathcal{X}_{i,k})$ , where  $\eta > 0$  is the learning rate,  $\nabla l(\cdot, \cdot)$  is the gradient of the loss function  $l(\cdot, \cdot)$ , and  $\mathcal{X}_{i,k}$  is a minibatch of the database (a subset of the local database  $\mathcal{D}_i$ ) in the  $k^{\text{th}}$  iteration.

At every round, each client initializes a gradient descent optimization algorithm using the latest received  $w^t$  and updates  $w_i^{t+1}$  via  $K$  iterations of the optimizer (4), i.e.,  $w_i^{t+1} = w_{i,K}$ . Optimal local parameters,  $w_i^{t+1}$ , are sent to the server for aggregation. After a sufficient number of global iterations between clients and the server (in the global counter  $t$ ) and local updates (in the local counter  $k$ ), the standard FL scheme converges to the optimal global model (3) (see [1]).

## 2.2 Privacy Requirements

Information about clients' private data can be inferred from the model updates throughout the training process [4, 5, 7, 8]. In addition, information leakages can also occur in the broadcasting step by analyzing the global model parameters [5]. In FL, two types of actors can infer private information: internal actors (participating clients, the central server, and third parties) and external actors (model consumers and eavesdroppers) [9]. We assume all the internal actors are untrusted (honest-but-curious), which means they will faithfully follow the designed FL protocol but attempt to infer private information. External actors are also untrusted; they aim to eavesdrop the communication between internal actors to infer information.

To address the problem of deducing private information of clients from their uploaded local models, numerous cryptographic privacy-preserving methods such as MPC and HE and perturbation-based techniques such as DP are usually employed to ensure that clients' local models are not accessed by the server or any malicious actors. Generally, these methods are designed to ensure that local models are not directly exposed to other parties to prevent inferring sensitive data.

Unfortunately, approaches exclusively using cryptographic methods remain vulnerable to inference over the aggregated models. As the aggregated models in every iteration remain unchanged from function execu-

tion without privacy, it has been shown that the server can recover local models through multiple intermediate global models [6]. Therefore, we must also consider potential inference over the aggregated models. Solutions addressing the aggregated models' privacy mostly use a DP framework. However, there is a trade-off between DP and the performance of FL, both in terms of model accuracy and convergence rate due to the added noises. In this study, we concentrate on protecting privacy against inference over intermediate local and global models, in the aggregation and broadcasting steps of FL without degrading the accuracy and convergence rate.

## 2.3 Privacy-preserving FL Problem

To prevent inference of the clients' databases from their local updates, we propose a privacy-preserving FL scheme to distort local updates  $w_i^t$  before transmission. Starting from the initial local model of clients  $w_{i,0}$ , which is the latest broadcasted global model by the server,  $w_{i,0} = w^t$ , we let the server distort the original aggregated update before disclosure through some encoding map  $\pi_1 : \mathbb{R}^n \rightarrow \mathbb{R}^{\tilde{n}}$ ,  $\tilde{w}^t = \pi_1(w^t)$ , and send the distorted  $\tilde{w}^t$  to the clients to run the optimization algorithm and update their local models. In general, running the standard optimization algorithm (4) on the distorted initial model  $\tilde{w}_{i,0} = \tilde{w}^t$  will not yield the same model update  $w_{i,K}$  that would be obtained if it was run using  $w^t$ . That is privacy-preserving FL methods that do not account for (remove) the distortion induced by the encoding map  $\pi_1(\cdot)$  lead to performance degradation. This is precisely the problem with perturbation-based techniques for privacy preservation (like with standard, non-homomorphic DP tools).

To address these challenges, the scheme proposed here seeks to design a new gradient descent optimization algorithm (referred hereafter to as the target optimizer) that runs on encoded model from the server  $\tilde{w}^t$  and returns an encoded local model update,  $\tilde{w}_i^{t+1}$ , that can be later decoded after aggregating local updates at the server side. Note that, we are not looking for one particular target algorithm but a methodology that can construct such target algorithm for any (or a broad class of) original optimization algorithm.

We seek to design  $\pi_1(\cdot)$  such that  $\tilde{w}^t = \pi_1(w^t) \in \mathbb{R}^{\tilde{n}}$  is of higher dimension than  $w^t \in \mathbb{R}^n$ , i.e.,  $\tilde{n} > n$ . We impose this condition to create redundancy in both the encoding map and target optimizer. Redundancy will allow us to inject randomness that can be traced through the algorithm, removed after model aggregation, and used to enforce an arbitrary level of differential privacy. Consider the higher-dimensional target optimizer:

$$\text{Target optimizer} \begin{cases} \tilde{w}_{i,0} = \tilde{w}^t, \\ \tilde{w}_{i,k+1} = \tilde{f}(\tilde{w}_{i,k}, \mathcal{D}_i), \\ k = 0, 1, \dots, K-1, \\ \tilde{w}_i^{t+1} = \tilde{w}_{i,K}, \end{cases} \quad (5)$$

with function  $\tilde{f} : \mathbb{R}^{\tilde{n}} \rightarrow \mathbb{R}^{\tilde{n}}$ ,  $\tilde{n}_y > n_y$ , to be designed, distorted initial model parameters  $\tilde{w}^t$  (the latest encoded global update from the server), and distorted local update  $\tilde{w}_i^{t+1}$  generated by the target optimizer.

Our goal is to design the encoding map  $\pi_1(\cdot)$  and the functions  $\tilde{f}(\cdot)$  such that the target optimizer can work on the encoded data  $\tilde{w}_{i,0} = \tilde{w}^t$  to produce encoded model update  $\tilde{w}_i^{t+1}$  that can be used to extract  $w^{t+1}$  after aggregation. Note that the choice of  $\tilde{f}(\cdot)$  in (5) provides a prescriptive design in terms of the standard optimization function  $f(\cdot)$  in (4).

In our setting, we consider a third party other than the clients and the server for model aggregation. We refer to this party simply as the *aggregator*. We consider that once a complete cycle has been finished at every client by the target optimizer, all clients send their last iteration,  $\tilde{w}_i^{t+1} = \tilde{w}_{i,K}$ , to the aggregator for model aggregation. The role of the aggregator is to interface between clients and the server and thus prevent the server from accessing exact encoded local models since it has access to the encoding map  $\pi_1(\cdot)$ . The aggregator takes the updated encoded local models from all clients,  $\tilde{w}_i^{t+1}$ , aggregates them, and sends (only) the aggregated encoded model to the server. Hence, the server cannot access any local model and only has access to the aggregated results. Since the aggregator only accesses the encoded local updates  $\tilde{w}_i^{t+1}$  and does not have access to  $\pi_1(\cdot)$ , it is not required to be trusted.

At a system-theoretic level, what we seek to accomplish is to embed model parameters  $w_{i,k}$  of the standard optimizer (initialized with the latest model from the server  $w^t$ ) into model parameters  $\tilde{w}_{i,k}$  of the target optimizer (initialized with the latest encoded model from the server  $\tilde{w}^t$ ). That is, we aim to design  $\pi_1(\cdot)$  and the target optimizer so that there exists a bijection between model parameters of both optimizers (referred here to as the *immersion map*), and thus, having model parameters of the target optimizer uniquely determining the model parameters of the standard one through the immersion map. This leads to the possibility of running the target optimizer (instead of the standard optimizer) by clients on encoded parameters  $\tilde{w}^t$ , and then, aggregating their local updates  $\tilde{w}_i^{t+1}$  by the aggregator to achieve aggregated encoded model  $\tilde{w}^{t+1}$ , which can be used to extract the exact aggregated global model  $w^{t+1}$  from  $\tilde{w}^{t+1}$  by the server. Hence, a fundamental question is how do we design  $\pi_1(\cdot)$  and the target optimizer  $\tilde{f}(\cdot)$  in (5) to accomplish this bijection?

In system and control theory, this type of embedding between systems trajectories is referred to as *system immersion* and has been used for nonlinear adaptive control [27, 33] and output regulation [34, 35]. Using system immersion tools in the context of privacy in cloud computing has been studied in [28]. In what follows, we explore the idea of using system immersion tools to design a privacy-preserving federated learning algorithm. We develop the necessary mathematical machinery and provide sufficient conditions to simultaneously design the

encoding map  $\pi_1(\cdot)$  and the target optimizer  $\tilde{f}(\cdot)$  to accomplish immersion and aggregated model extraction using ideas from system immersion. This will culminate in a problem description on immersion-based coding for privacy-preserving FL at the end of this section.

#### 2.4 Immersion-based Privacy-preserving FL: Secure Aggregation

Consider the standard and target optimization algorithms in (4) and (5), respectively. We say that standard optimizer is immersed in target optimizer if  $\tilde{n} > n$  and there exists a function  $\pi_1 : \mathbb{R}^n \rightarrow \mathbb{R}^{\tilde{n}}$  that satisfies  $\tilde{w}_{i,k} = \pi_1(w_{i,k})$  for all  $w_{i,k}$  and  $\tilde{w}_{i,k}$  generated by the optimizer and target optimizer, respectively. That is, any model parameters of the target optimizer are model parameters of the original optimizer through the mapping  $\pi_1(\cdot)$ , and  $\pi_1(\cdot)$  is an immersion because the dimension of its image is  $\tilde{n} > n$ . We refer to this map  $\pi_1(\cdot)$  as the *immersion map*.

To guarantee that the standard optimizer (4) is immersed in the target optimizer (5) (in the sense introduced above), we need to impose conditions on the functions shaping the optimizers, their initial conditions, and the immersion map, i.e., on  $(f, w_{i,0}, \pi_1, \tilde{f}, \tilde{w}_{i,0})$ . In particular, we require to design  $(\pi_1, \tilde{f}, \tilde{w}_{i,0})$  such that  $\tilde{w}_{i,k} = \pi_1(w_{i,k})$  for all  $k$ , i.e., the manifold  $\tilde{w}_{i,k} = \pi_1(w_{i,k})$  must be forward invariant under the optimization algorithms in (4) and (5) [36]. Let us define the off-the-manifold error  $e_k := \tilde{w}_{i,k} - \pi_1(w_{i,k})$ . The manifold  $\tilde{w}_{i,k} = \pi_1(w_{i,k})$  is forward invariant if and only if the origin of the error dynamics:

$$\begin{aligned} e_{k+1} &= \tilde{w}_{i,k+1} - \pi_1(w_{i,k+1}) \\ &= \tilde{f}(e_k + \pi_1(w_{i,k}), \mathcal{D}_i) - \pi_1(f(w_{i,k}, \mathcal{D}_i)), \end{aligned} \quad (6)$$

is a fixed point, i.e.,  $e_k = \mathbf{0}$  implies  $e_{k+1} = \mathbf{0}$  for all  $k \in \mathbb{N}_0$  [37]. Substituting  $e_k = \mathbf{0}$  and  $e_{k+1} = \mathbf{0}$  in (6) leads to

$$\tilde{f}(\pi_1(w_{i,k}), \mathcal{D}_i) - \pi_1(f(w_{i,k}, \mathcal{D}_i)) = \mathbf{0}. \quad (7)$$

Therefore,  $\tilde{w}_{i,k} = \pi_1(w_{i,k})$  is satisfied for all  $k$  if: **(1)** the initial condition of (5),  $\tilde{w}_{i,0}$ , satisfies  $\tilde{w}_{i,0} = \pi_1(w_{i,0})$ , which leads to  $e_0 = \mathbf{0}$  (start on the manifold); and **(2)** the dynamics of both algorithms match under the immersion map, i.e., (7) is satisfied (invariance condition on the manifold) for all  $k \geq 0$ . We refer to these two conditions as the *immersion conditions*.

---

#### Immersion Conditions:

$$\begin{cases} \tilde{f}(\pi_1(w_{i,k}), \mathcal{D}_i) = \pi_1(f(w_{i,k}, \mathcal{D}_i)), & (\text{invariance}) \\ \tilde{w}_{i,0} = \pi_1(w_{i,0}). & (\text{start on the manifold}) \end{cases} \quad (8)$$

---

The start on the manifold condition is satisfied by encod-

ing the aggregated global model by the server through the mapping  $\pi_1(\cdot)$  at every global iteration as  $\tilde{w}^t = \pi_1(w^t)$ . The server broadcasts the encoded aggregated model  $\tilde{w}^t$  to all clients. Then, each client initializes their target optimizer using the latest received encoded global model as  $\tilde{w}_{i,0} = \pi_1(w_{i,0}) = \tilde{w}^t$  and updates  $\tilde{w}_i^{t+1}$  via  $K$  iterations of the target optimizer (5), i.e.,  $\tilde{w}_i^{t+1} = \tilde{w}_{i,K}$ . So far, we have derived sufficient conditions (8) for local model parameters of the standard optimizer to be immersed into the model parameters of the target optimizer in terms of  $(f, \tilde{f}, w_{i,0}, \tilde{w}_{i,0}, \pi_1)$ . Next, we derive conditions on the immersion map  $\pi_1(\cdot)$  so that the encoded aggregated model by the aggregator,  $\tilde{w}_a^{t+1}$ , can be decoded to extract the original aggregated model  $w^{t+1}$  by the server. The aggregated model by the aggregator at iteration  $t$  is given by

$$\tilde{w}_a^{t+1} = \sum_{i=1}^{N_c} \frac{|\mathcal{D}_i|}{|\mathcal{D}|} \tilde{w}_i^{t+1} = \sum_{i=1}^{N_c} \frac{|\mathcal{D}_i|}{|\mathcal{D}|} \pi_1(w_i^{t+1}), \quad (9)$$

where the right-hand side part of (9) follows from the immersion condition  $\tilde{w}_{i,K} = \pi_1(w_{i,K})$ . The server receives  $\tilde{w}_a^{t+1}$  in (9) and aims to retrieve  $w^{t+1} = \sum_{i=1}^{N_c} (|\mathcal{D}_i|/|\mathcal{D}|) w_i^k$  – the aggregated result of the standard optimization algorithm in (4). The latter imposes an extra condition on the immersion map,  $\pi_1(\cdot)$ , since to retrieve  $w^{t+1}$  from  $\tilde{w}^{t+1}$ , there must exist a left-inverse function  $\pi_1^L : \mathbb{R}^{\tilde{n}} \rightarrow \mathbb{R}^n$  of  $\pi_1(\cdot)$ , i.e., satisfying the following left-invertibility condition  $\pi_1^L(\tilde{w}^{t+1}) = w^{t+1}$ :

$$\pi_1^L(\tilde{w}^{t+1}) = \pi_1^L\left(\sum_{i=1}^{N_c} \frac{|\mathcal{D}_i|}{|\mathcal{D}|} \pi_1(w_i^{t+1})\right) = \sum_{i=1}^{N_c} \frac{|\mathcal{D}_i|}{|\mathcal{D}|} w_i^{t+1}. \quad (10)$$

If such  $\pi_1^L(\cdot)$  and  $\pi_1(\cdot)$  exist, the server can retrieve the original aggregated parameters  $w^{t+1} = \sum_{i=1}^{N_c} (|\mathcal{D}_i|/|\mathcal{D}|) w_i^{t+1}$  by passing the encoded results through the function  $\pi_1^L(\cdot)$ . We now have all the machinery to state the problem we seek to solve.

**Problem 1 (Immersion-based Privacy-Preserving FL)** For given  $(f, w_{i,0})$  of the standard optimizer (4), design an immersion map  $\pi_1(\cdot)$ , and  $(\tilde{f}, \tilde{w}_{i,0})$  of the target optimizer (5) so that: **(a)** the immersion conditions (8) hold; and **(b)** there exists a function  $\pi_1^L(\cdot)$  satisfying (10).

**Remark 1** Solutions to Problem 1 characterize a class of encoding maps and target optimizers for which we can design homomorphic encryption schemes (a prescriptive design for given  $(f, w_{i,0})$ ). However, this class is infinite-dimensional (over a function space). It leads to an underdetermined algebraic problem with an infinite-dimensional solution space. To address this aspect, we impose structure on the maps we seek to design. We restrict to random affine maps composed of linear coordinate transformations and additive random processes. In

what follows, we prove that this class of maps is sufficient to guarantee an arbitrary level of differential privacy.

### 3 Immersion-based Coding for Privacy-Preserving FL

In this section, we construct a prescriptive solution to Problem 1 using a random affine immersion map  $\pi_1(\cdot)$ . As the problem formulation and solution are based on the system immersion theory, we refer to our algorithm as *System Immersion based Federated Learning* (SIFL). Let the immersion map  $\pi_1(\cdot)$  be an affine function of the form:

$$\pi_1(w_{i,k}) := \Pi_1 w_{i,k} + b_1^t, \quad (11)$$

for matrix  $\Pi_1 \in \mathbb{R}^{\tilde{n} \times n}$ ,  $\tilde{n} > n$ , to be designed, and some i.i.d. multivariate random process  $b_1^t \in \mathbb{R}^{\tilde{n}}$ . For this map, the immersion condition (8) amounts to  $\tilde{w}_{i,0} = \pi_1(w_{i,0})$  (encoded by the server before broadcasting) and

$$\tilde{f}(\Pi_1 w_{i,k} + b_1^t, \mathcal{D}_i) = \Pi_1 f((w_{i,k}), \mathcal{D}_i) + b_1^t. \quad (12)$$

Let the function  $\tilde{f}(\cdot)$  be designed in the following form (using the form of the original optimizer in (4)):

$$\tilde{f}(\tilde{w}_{i,k}, \mathcal{D}_i) := \tilde{w}_{i,k} - M_2 g(M_1 \tilde{w}_{i,k}, \mathcal{D}_i), \quad (13)$$

for some matrices  $M_1 \in \mathbb{R}^{n \times \tilde{n}}$  and  $M_2 \in \mathbb{R}^{\tilde{n} \times n}$  to be designed. Hence, the immersion condition (12) takes the form:

$$\begin{aligned} \Pi_1 w_{i,k} + b_1^t - M_2 g(M_1 (\Pi_1 w_{i,k} + b_1^t), \mathcal{D}_i) \\ = \Pi_1 (w_{i,k} - g(w_{i,k}, \mathcal{D}_i)) + b_1^t. \end{aligned} \quad (14)$$

Note that the choice of  $\tilde{f}(\cdot)$  in (13) provides a prescriptive design in terms of the standard optimization function  $f(\cdot)$  in (4). That is, we exploit the knowledge of the original optimizer and build the target optimizer on top of it in an algebraic manner.

To satisfy (14), we must enforce  $M_1(\Pi_1 w_{i,k} + b_1^t) = w_{i,k}$  and  $M_2 = \Pi_1$ , which implies that  $M_1 \Pi_1 = I$ ,  $M_1 b_1^t = \mathbf{0}$  (i.e.,  $b_1^t \in \ker[M_1]$ ). From this brief analysis, we can draw the following conclusions: 1)  $\Pi_1$  must be of full column rank (i.e.,  $\text{rank}[\Pi_1] = n$ ); 2)  $M_1$  is the left inverse of  $\Pi_1$ , i.e.,  $M_1 = \Pi_1^L$  (which always exists given the rank of  $\Pi_1$ ); 3)  $b_1^t \in \ker[\Pi_1^L]$  (this kernel is always nontrivial because  $\Pi_1$  is full column rank by construction); and 4)  $M_2 = \Pi_1$ . Combining all these facts, the final form of  $\tilde{f}(\cdot)$  in (5) is given by

$$\tilde{f}(\tilde{w}_{i,k}, \mathcal{D}_i) := \tilde{w}_{i,k} - \Pi_1 g(\Pi_1^L \tilde{w}_{i,k}, \mathcal{D}_i). \quad (15)$$

At every global iteration  $t$ , vector  $b_1^t$  is designed by the server to satisfy  $\Pi_1^L b_1^t = \mathbf{0}$  and used to distort the coding map  $\pi_1(\cdot)$  in (11). The role of  $b_1^t$  is to randomize the model parameters so that we can guarantee differential privacy. The idea is that because the server

(that applies mapping  $\pi_1(\cdot)$  to the global model before sending it to clients) knows that  $b_1^t$  draws realizations from the kernel of  $\Pi_1^L$ , the distortion induced by it can be removed at the server side after aggregation by the aggregator. To enforce that  $b_1^t \in \ker[\Pi_1^L]$ , without loss of generality, we let it be of the form  $b_1^t = N_1 r_1^t$  for some matrix  $N_1 \in \mathbb{R}^{\tilde{n} \times (\tilde{n}-n)}$  expanding the kernel of  $\Pi_1^L$  (i.e.,  $\Pi_1^L N_1 = \mathbf{0}$ ) and an arbitrary i.i.d. process  $r_1^t \in \mathbb{R}^{(\tilde{n}-n)}$ . This structure for  $b_1^t$  always satisfies  $\Pi_1^L b_1^t = \Pi_1^L N_1 r_1^t = \mathbf{0}$ , for  $r_1^t$  with arbitrary probability distribution.

So far, we have designed  $(\tilde{f}(\cdot), \pi_1(\cdot))$  to satisfy the immersion conditions (8) for the affine maps in (11). Next, we design the extracting function  $\pi_1^L(\cdot)$  satisfying the left invertibility condition (10), that extracts the true global model  $w^{t+1}$  from the encoded  $\tilde{w}^{t+1}$ . Substituting the designed immersion map (11) in the aggregated encoded model (9) yields

$$\begin{aligned} \tilde{w}_a^{t+1} &= \sum_{i=1}^{N_c} \frac{|\mathcal{D}_i|}{|\mathcal{D}|} (\Pi_1 w_{i,K} + b_1^t) \\ &= \Pi_1 \left( \sum_{i=1}^{N_c} \frac{|\mathcal{D}_i|}{|\mathcal{D}|} w_{i,K} \right) + b_1^t = \Pi_1 w^{t+1} + b_1^t, \end{aligned} \quad (16)$$

where  $\tilde{w}_a^{t+1}$  and  $w^{t+1}$  denote the encoded and original aggregated models, respectively. Given (16), condition (10) can be written as

$$\pi_1^L(\tilde{w}_a^{t+1}) = \pi_1^L(\Pi_1 w^{t+1} + b_1^t) = w^{t+1}, \quad (17)$$

which trivially leads to

$$\pi_1^L(\tilde{w}_a^{t+1}) := \Pi_1^L \tilde{w}_a^{t+1}, \quad (18)$$

since  $\Pi_1^L \Pi_1 = I$  and  $b_1^t \in \ker[\Pi_1^L]$ .

We can now state the proposed solution to Problem 1. **Proposition 1 (Solution to Problem 1)** *For given full rank matrix  $\Pi_1 \in \mathbb{R}^{\tilde{n} \times n}$ , matrix  $N_1 \in \mathbb{R}^{\tilde{n} \times (\tilde{n}-n)}$  expanding the kernel of  $\Pi_1^L$  (i.e.,  $\Pi_1^L N_1 = \mathbf{0}$ ), and random process  $r_1^t \in \mathbb{R}^{(\tilde{n}-n)}$ , the encoding map:*

$$\tilde{w}^t := \Pi_1 w^t + N_1 r_1^t, \quad (19)$$

target optimizer:

$$\begin{cases} \tilde{w}_{i,0} = \tilde{w}^t, \\ \tilde{w}_{i,k+1} = \tilde{f}(\tilde{w}_{i,k}, \mathcal{D}_i) := \tilde{w}_{i,k} - \Pi_1 g(\Pi_1^L \tilde{w}_{i,k}, \mathcal{D}_i), \\ k = 0, 1, \dots, K-1, \\ \tilde{w}_i^{t+1} = \tilde{w}_{i,K}, \end{cases} \quad (20)$$

and inverse function:

$$\pi_1^L(\tilde{w}^{t+1}) := \Pi_1^L \tilde{w}^{t+1}, \quad (21)$$

provide a solution to Problem 1.

**Proof:** Proposition 1 follows from the discussion provided in this section above. ■

### 3.1 Summary SIFL Algorithm solving Problem 1: Privacy-Preserving Aggregation

The summary of the algorithm is as follows:

- **FL initialization and encoding by the server.** The server initializes the global model  $w^0$  and encodes it as  $\tilde{w}^0 = \Pi_1 w^0 + b_1^0$ . Then, it immerses the standard optimizer into the target optimization algorithm as in (20) and broadcasts  $\tilde{w}^0$ , the target optimizer  $\tilde{f}(\cdot)$ , and other hyperparameters to clients.
- **Local model training and update by clients.** The clients receive the current encoded global model  $\tilde{w}^t$  sent by the server and update their local model parameters using their local databases  $\mathcal{D}_i$  and the target optimizer (20). Then, clients send their encoded local updates  $\tilde{w}_i^{t+1}$  to the aggregator for aggregation.
- **Global model aggregation.** The aggregator takes the average of local encoded models and sends the aggregated model  $\tilde{w}_a^{t+1}$  in (16) to the server.
- **Global model decoding and encoding and broadcasting by the server.** The server decodes the aggregated global model using the inverse function  $\pi_1^L(\cdot)$  in (21). Then, it encodes the new global model using the immersion map  $\pi_1(\cdot)$  (19) and broadcasts it to all clients for the next round.

The pseudo-code of SIFL for privacy-preserving model aggregation is shown in Algorithm 1.

## 4 Global model privacy in FL

In the previous section, we present the immersion-based coding scheme to provide privacy for local and global models of FL. Using this scheme, none of the internal and external actors can access the original local and global models; only the server can access the exact aggregated models in each iteration. It has been shown that the server can recover local models by accessing multiple intermediate global models [6]. Therefore, we next consider potential inference over the aggregated models at the server side.

In this section, we provide a similar privacy-preserving mechanism based on an additional encoding of the aggregated model by the aggregator, using a mapping  $\pi_2(\cdot)$ , and decoding it by clients on the next iteration, to avoid server access to intermediate global models during the training process. Let  $\tilde{w}^t$ ,  $\bar{w}^t$ , and  $w^{t'}$  denote the global models encoded using  $\pi_1(\cdot)$  by the server, global models encoded using  $\pi_2(\cdot)$  by the aggregator, and global models encoded using both  $\pi_1(\cdot)$  and  $\pi_2(\cdot)$ , respectively. In this scheme, the aggregator encodes

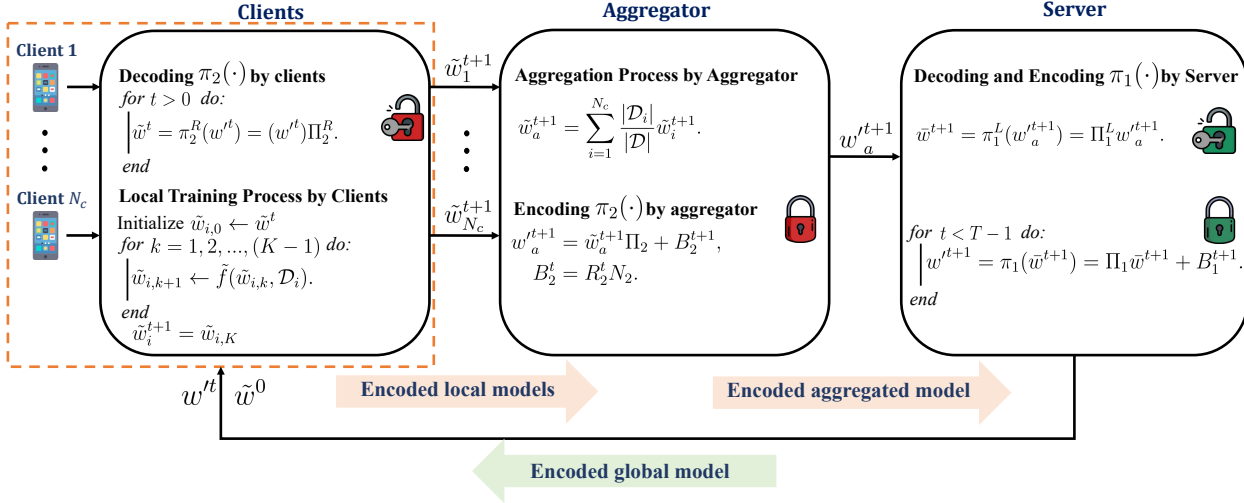


Fig. 1. Flowchart of the extended SIFL method.

**Algorithm 1** SIFL algorithm solving Problem 1: Privacy-Preserving Aggregation

**Input:** Clients  $\mathcal{C}_i$  and their databases  $\mathcal{D}_i$ ,  $i \in \{1, 2, \dots, N_c\}$ , FL iterations  $T$ , hyperparameters of the specific gradient-descent optimizer, local iterations  $K$ , immersion mapping matrix  $\Pi_1 \in \mathbb{R}^{\tilde{n} \times n}$ , its left inverse  $\Pi_1^L$ , and matrix  $N_1 \in \ker[\Pi_1^L]$ .

**Output:** Trained global model  $w^T$ .

**Handshaking phase:**

The server sends target optimizer  $\tilde{f}$  as in (20), the encoded initialized global model  $\tilde{w}^0$ , and other hyperparameters to clients for model update.

**for** each global iteration  $t = 0, 1, \dots, T - 1$  **do**

*Local Training Process by Clients:*

**for** each  $\mathcal{C}_i$  **do**

Initialize:  $\tilde{w}_{i,0} \leftarrow \tilde{w}^t$ .

**for** each local iteration  $k = 1, 2, \dots, (K - 1)$  **do**

$\tilde{w}_{i,k+1} \leftarrow \tilde{f}(\tilde{w}_{i,k}, \mathcal{D}_i)$ .

Clients send  $\tilde{w}_i^{t+1} = \tilde{w}_{i,K}$  to aggregator.

**end for**

**end for**

*Aggregation Process by Aggregator:*

$$\tilde{w}_a^{t+1} = \sum_{i=1}^{N_c} \frac{|\mathcal{D}_i|}{|\mathcal{D}|} \tilde{w}_i^{t+1}.$$

The aggregator sends  $\tilde{w}_a^{t+1}$  to the server.

*Decoding and Encoding Process by Server:*

$w^{t+1} = \pi_1^L(\tilde{w}_a^{t+1})$ .

**for**  $t < T - 1$  **do**

$\tilde{w}^{t+1} = \pi_1(w^{t+1})$ .

**end for**

Server sends encoded  $\tilde{w}^{t+1}$  to clients.

**end for**

the aggregated model using a right invertible encoding map  $\pi_2(\cdot)$  and sends the encoded aggregated model

$w_a'^{t+1} = \pi_2(\tilde{w}_a^{t+1})$  to the server. Then, in the following iteration, clients decode the encoded aggregated model using the right inverse of the mapping  $\pi_2(\cdot)$ , denoted as  $\pi_2^R(\cdot)$ , i.e.,  $\pi_2 \circ \pi_2^R(s) = s$ , and use the encoded model (which is still encoded by the immersion map  $\pi_1(\cdot)$  of the server) for the model update. Then, the server needs to be able to apply its decoding and encoding maps,  $\pi_1^L(\cdot)$  and  $\pi_1(\cdot)$ , to the encoded model by the aggregator  $w_a'^{t+1} = \pi_2(\tilde{w}_a^{t+1})$  without decoding  $\pi_2(\cdot)$ . Furthermore, clients need to be able to decode  $\pi_2(\cdot)$  using  $\pi_2^R(\cdot)$  after encoding by the server to employ it as the initial condition for the target optimizer. Therefore, coding  $\pi_2(\cdot)$  is also a homomorphic encryption scheme, meaning that the server needs to be able to apply its coding on the encoded  $w_a'^{t+1}$  without decoding  $\pi_2(\cdot)$  and then, the clients need to be able to extract  $\tilde{w}^{t+1}$  from the encoded  $w^{t+1}$  using the right inverse function  $\pi_2^R(\cdot)$ . These conditions can be formulated as follows.

The server receives the encoded aggregated model  $w_a'^{t+1} = \pi_2(\tilde{w}_a^{t+1})$  and decode mapping  $\pi_1(\cdot)$  as follows:

$$\bar{w}^{t+1} = \pi_1^L(w_a'^{t+1}). \quad (22)$$

Then, server encodes  $\bar{w}^{t+1}$  using mapping  $\pi_1(\cdot)$  before broadcasting it to clients as  $w^{t+1} = \pi_1(\bar{w}^{t+1})$ . Clients receive  $w^{t+1}$  and aim to retrieve  $\tilde{w}^{t+1}$ , which is still encoded using the immersion map of the server  $\pi_1(\cdot)$ , to employ it as the initial condition for the target optimizer. Hence, there must exist a right inverse function  $\pi_2^R(\cdot)$  satisfying the following right invertibility condition:

$$\tilde{w}^{t+1} = \pi_2^R(w^{t+1}). \quad (23)$$

Note that the aggregator needs to share the decoding map  $\pi_2^R(\cdot)$  with clients in the handshaking phase. With this approach, the server does not have access to any



local or intermediate global models. The problem of designing the aggregator mapping can be defined as follows.

**Problem 2 (Coding for Global Model Privacy)** For given immersion map of the server  $\pi_1(\cdot)$  (19) and its inverse  $\pi_1^L(\cdot)$  (21), design an immersion map  $\pi_2(\cdot)$  so that: **(a)** mapping  $\pi_1(\cdot)$  can be decoded after applying mapping  $\pi_2(\cdot)$  to the aggregated model using  $\pi_1^L(\cdot)$ , i.e., (22); and **(b)** there exists a function  $\pi_2^R(\cdot)$  satisfying (23).

#### 4.1 Affine Solution to Problem 2

As we proposed in the solution to Problem 1, let the encoding map  $\pi_2(\cdot)$  be an affine function of the form:

$$w_a^{t+1} = \pi_2(\tilde{w}_a^{t+1}) := \tilde{w}_a^{t+1}\Pi_2 + B_2^{t+1}, \quad (24)$$

for vector  $\Pi_2 \in \mathbb{R}^{1 \times p}$ ,  $p > 1$ , and matrix  $B_2^t \in \mathbb{R}^{\tilde{n} \times p}$ . We design  $\pi_2(\cdot)$  such that  $w_a^t = \pi_2(\tilde{w}_a^t) \in \mathbb{R}^{\tilde{n} \times p}$  is of higher dimension than  $\tilde{w}_a^t \in \mathbb{R}^{\tilde{n}}$ . We impose this design to create redundancy in the encoding map, which allows us to inject randomness that can be 1) traced through the coding algorithm at the server side, 2) be removed after at the clients' side, and 3) enforce an arbitrary level of DP for the aggregated models. The reason why the increase in the dimension of the aggregated model  $\tilde{w}_a^t$  is from the right side (rather than similar to the previous mapping  $\pi_1(\cdot)$  where we increase the dimension from the left side as shown in (11)) is to be able to decode mapping  $\pi_1(\cdot)$ , with the left inverse map  $\pi_1^L(\cdot)$  at the server (22), and to decode mapping  $\pi_2(\cdot)$  with the right inverse map  $\pi_2^R(\cdot)$  at clients (23).

By substituting (24) in the invertibility condition of mapping  $\pi_1(\cdot)$  (22), we have:

$$\tilde{w}^{t+1} = \pi_1^L(\tilde{w}_a^{t+1}\Pi_2 + B_2^{t+1}) = w^{t+1}\Pi_2 + \Pi_1^L B_2^{t+1}. \quad (25)$$

Hence, using the proposed structure of mapping  $\pi_2(\cdot)$ , mapping  $\pi_1(\cdot)$  can be still decoded using  $\pi_1^L(\cdot)$ . Note that, to be able to apply the encoding map  $\pi_1(\cdot)$  to  $\tilde{w}^t$ , according to its dimension  $\tilde{w}^t \in \mathbb{R}^{\tilde{n} \times p}$ , the dimension of mapping  $\pi_1(\cdot)$  must be set to  $\pi_1: \mathbb{R}^{\tilde{n} \times p} \rightarrow \mathbb{R}^{\tilde{n} \times p}$  (rather than the previous design in equation (11),  $\pi_1: \mathbb{R}^n \rightarrow \mathbb{R}^{\tilde{n}}$ ). Hence,  $\pi_1(\cdot)$  with this dimension is as follows:

$$\pi_1(\tilde{w}^t) = \Pi_1 \tilde{w}^t + B_1^t, \quad (26)$$

where the additive random term  $B_1^t$  should be designed with dimension  $B_1^t = N_1 R_1^t \in \mathbb{R}^{\tilde{n} \times p}$ , which can be achieved by setting the dimension of the random term as  $R_1^t \in \mathbb{R}^{(\tilde{n}-n) \times p}$ . Hence, the encoded model parameters after applying mapping  $\pi_1(\cdot)$  by the server is of the following form:

$$w^{t+1} = \pi_1(\tilde{w}^{t+1}) = \Pi_1 w^{t+1} \Pi_2 + \Pi_1 \Pi_1^L B_2^{t+1} + B_1^{t+1}. \quad (27)$$

Clients receive encoded  $w^t$  and aim to decode the aggregator coding  $\pi_2(\cdot)$  using the decoding map  $\pi_2^R(\cdot)$ , before updating the local models by the target optimizer (20). Substituting (27) in the right invertibility condition (23), the design condition for  $\pi_2^R(\cdot)$  amounts to:

$$\pi_2^R(w^{t+1}) = \pi_2^R(\Pi_1 w^t \Pi_2 + \Pi_1 \Pi_1^L B_2^t + B_1^t) = \Pi_1 w^t + b_1^t. \quad (28)$$

Let the decoding map be of the form  $\pi_2^R(w^{t+1}) = w^{t+1} M_3$  with  $M_3 \in \mathbb{R}^{p \times 1}$ . Hence, condition (28) takes the form

$$(\Pi_1 w^t \Pi_2 + \Pi_1 \Pi_1^L B_2^t + B_1^t) M_3 = \Pi_1 w^t + b_1^t. \quad (29)$$

Hence, to satisfy (29), we must have  $\Pi_2 M_3 = I$  and  $B_2^t M_3 = \mathbf{0}$ . Moreover, the vector  $b_1^t = B_1^t M_3 \in \mathbb{R}^{\tilde{n} \times 1}$  should work as the additive random term in  $\pi_1(\cdot)$  satisfying the necessary condition of  $b_1^t \in \ker[\Pi_1^L]$  ( $\Pi_1^L b_1^t = 0$ ). It follows that: 1) vector  $M_3$  is a right inverse of  $\Pi_2$ , i.e.,  $M_3 = \Pi_2^R$ ; 2)  $B_2^t$  is in the right null space of  $\Pi_2^R$  ( $B_2^t \in \ker[\Pi_2^R]$ ); this null space is always nontrivial because  $\Pi_2$  is a row vector. Then,  $b_1^t := B_1^t \Pi_2^R$  always satisfy the condition  $\Pi_1^L b_1^t = 0$  provided that  $\Pi_1^L B_1^t = \mathbf{0}$ . The final form for  $\pi_2^R(\cdot)$  is given by

$$\pi_2^R(w^{t+1}) := w^{t+1} \Pi_2^R. \quad (30)$$

At every global iteration  $t$ , the aggregator designs a matrix  $B_2^t$  satisfying  $B_2^t \Pi_2^R = \mathbf{0}$  and uses it to construct the encoding map  $\pi_2(\tilde{w}^t) = \tilde{w}^t \Pi_2 + B_2^t$  (the encoding scheme). The role of  $B_2^t$  is to randomize the aggregated model parameters so that we can guarantee DP for intermediate global models. Using the same reasoning as for the design of  $b_1^t$ , to enforce that  $B_2^t \in \ker[\Pi_2^R]$ , without loss of generality, we let it be of the form  $B_2^t = R_2^t N_2$  for some matrix  $N_2 \in \mathbb{R}^{(p-1) \times p}$  expanding the kernel of  $\Pi_2^R$  (i.e.,  $N_2 \Pi_2^R = \mathbf{0}$ ) and an arbitrary i.i.d. process  $R_2^t \in \mathbb{R}^{\tilde{n} \times (p-1)}$ . This structure for  $B_2^t$  always satisfies  $B_2^t \Pi_2^R = R_2^t N_2 \Pi_2^R = \mathbf{0}$ , for  $R_2^t$  with arbitrary probability distribution.

We can now state the proposed solution to Problem 2. **Proposition 2 (Solution to Problem 2)** For given vector  $\Pi_2 \in \mathbb{R}^{1 \times p}$ , matrix  $N_2 \in \mathbb{R}^{(p-1) \times p}$  expanding the kernel of  $\Pi_2^R$  (i.e.,  $N_2 \Pi_2^R = \mathbf{0}$ ), and random process  $R_2^t \in \mathbb{R}^{\tilde{n} \times (p-1)}$ , the encoding map:

$$w^t = \pi_2(\tilde{w}_a^t) := \tilde{w}_a^t \Pi_2 + R_2^t N_2, \quad (31)$$

and inverse function:

$$\pi_2^R(w^{t+1}) := w^{t+1} \Pi_2^R, \quad (32)$$

provide a solution to Problem 2.

**Proof:** Proposition 2 follows from the discussion provided above. ■

**Remark 2** In Algorithm 1 of the method, we present a cryptography system to provide privacy for local and

global models of FL in Proposition 1. Using this scheme, none of the internal and external actors have access to the original local and global models, except that the server only has access to the exact aggregated model in each iteration. In the extension of the SIFL method proposed in this section, we design a similar coding to encode the aggregated model by the aggregator to avoid server access to the intermediate global models. Hence, this extension can prevent all internal actors (the server, clients, and aggregator) and external actors from accessing the intermediate local and global models of FL.

#### 4.2 Extended SIFL Algorithm solving Problems 1 and 2: Privacy-Preserving Aggregation and Broadcasting

The summary of the extended algorithm is as follows:

- **FL initialization.** The server initializes the global model  $w^0$  and encodes it as  $\tilde{w}^0 = \Pi_1 w^0 + b_1^0$ . Then, it immerses the standard optimizer into the target optimization algorithm as in (20) and broadcasts  $\tilde{w}^0$ , the target optimizer  $\tilde{f}(\cdot)$ , and other hyperparameters to clients. The aggregator sends the decoding map  $\pi_2^R(\cdot)$  in (32) to clients.
- **Local model training and update by clients.** At the first round ( $t = 0$ ), clients use the initial model  $\tilde{w}^0$  and the target optimizer (20) to train their local models. After the first round, the clients receive the encoded global model  $w'^t$ , encoded both by the server and the aggregator. They first decode the aggregator mapping using the inverse function  $\pi_2^R(\cdot)$  in (32) to extract  $\tilde{w}^t$ . Then, they update their local model parameters using their local databases  $\mathcal{D}_i$  and the target optimizer (20) and send their encoded local updates  $\tilde{w}_i^{t+1}$  to the aggregator for aggregation.
- **Global model aggregation.** The aggregator takes the average of local encoded models, encodes the aggregated model  $\tilde{w}_a^{t+1}$  with  $\pi_2(\cdot)$  in (31), and sends the aggregated encoded model  $w'^{t+1}$  to the server.
- **Global model decoding, encoding, and broadcasting by the server.** The server decodes the aggregated global model using the inverse function  $\pi_1^L(\cdot)$  in (18). Then, it encodes the new global model using the immersion map  $\pi_1(\cdot)$  (19) and broadcasts it to all clients for the next round.

The pseudo-code of our extended SIFL scheme is shown in Algorithm 2, and its flowchart is depicted in Figure 1.

## 5 Privacy Guarantees

The private element we consider in the proposed scheme is privacy of the clients' local databases  $\mathcal{D}_i$ . In what follows, we focus on how to enforce differential privacy of the encoding mechanisms  $\pi_1(\cdot)$  and  $\pi_2(\cdot)$  by properly selecting the random processes  $R_1^t$  and  $R_2^t$  and the encoding matrices in the affine maps of the server  $\tilde{w}^t = \Pi_1 w^t + N_1 R_1^t$  and the aggregator  $w'_t = \tilde{w}^t \Pi_2 + R_2^t N_2$ .

---

### Algorithm 2 Extended SIFL Method: Privacy-Preserving Local and Global Models

---

**Input:** Set of clients  $\mathcal{C}_i$  and their databases  $\mathcal{D}_i$ , number of FL iterations  $T$ , hyperparameters of the specific gradient descent optimizer, number of local optimizer iterations  $K$ , immersion mapping matrix  $\Pi_1 \in \mathbb{R}^{\tilde{n} \times n}$ , its left inverse  $\Pi_1^L$ , matrix  $N_1 \in \ker[\Pi_1^L]$ , vector  $\Pi_2 \in \mathbb{R}^{1 \times p}$ , its right inverse  $\Pi_2^R$ , matrix  $N_2 \in \ker[\Pi_2^R]$ .

**Output:** Trained global model  $w^T$ .

**Handshaking phase:**

The server sends target optimizer  $\tilde{f}(\cdot)$  as in (20), the encoded initialized global model  $\tilde{w}^0$ , and other hyperparameters to clients for model update. The aggregator sends the decoding key  $\Pi_2^R$  to clients.

**for** each global iteration  $t = 0, 1, \dots, T - 1$  **do**

**for** each  $\mathcal{C}_i$  **do**

**Decoding**  $\pi_2(\cdot)$  **by Clients:**

**for**  $t > 0$  **do**

$$\tilde{w}^t = \pi_2^R(w'^t) = (w'^t) \Pi_2^R. \quad (33)$$

**end for**

**Local Training Process by Clients:**

    Initialize:  $\tilde{w}_{i,0} \leftarrow \tilde{w}^0$

**for** each local iteration  $k = 1, 2, \dots, (K - 1)$  **do**

$$\tilde{w}_{i,k+1} \leftarrow \tilde{f}(\tilde{w}_{i,k}, \mathcal{D}_i).$$

**end for**

$\mathcal{C}_i$  sends  $\tilde{w}_i^t$  to the aggregator.

**end for**

**Aggregation Process by Aggregator:**

$$\tilde{w}_a^{t+1} = \sum_{i=1}^{N_c} \frac{|\mathcal{D}_i|}{|\mathcal{D}|} \tilde{w}_i^{t+1}.$$

**Encoding**  $\pi_2(\cdot)$  **by Aggregator:**

$$w'^{t+1}_a = \tilde{w}_a^{t+1} \Pi_2 + B_2^{t+1}, \\ B_2^t = R_2^t N_2.$$

  The aggregator sends  $w'^{t+1}_a$  to the server.

**Decoding and encoding**  $\pi_1(\cdot)$  **by Server:**

$$\bar{w}^{t+1} = \pi_1^L(w'^{t+1}_a) = \Pi_1^L w'^{t+1}_a.$$

**for**  $t < T - 1$  **do**

$$w'^{t+1} = \pi_1(\bar{w}^{t+1}) = \Pi_1 \bar{w}^{t+1} + B_1^{t+1}.$$

**end for**

  Server sends encoded  $w'^{t+1}$  to clients.

**end for**

---

We provide a tailored solution to guarantee DP for the class of mechanisms that we consider in (19) and (31). In particular, we prove that the proposed scheme, with full-column rank encoding  $\Pi_1$  and vector  $\Pi_2$ , can provide any desired level of differential privacy without reducing the accuracy and performance of the original algorithm.

## 5.1 Differential Privacy

In the context of databases,  $(\epsilon, \delta)$ -Differential Privacy (DP) [38] was introduced as a probabilistic framework to quantify privacy of probabilistic maps. The constant  $\epsilon \geq 0$  quantifies how similar (different) are outputs of a mechanism on *adjacent* datasets, say  $\mathcal{D}$  and  $\mathcal{D}'$ , and  $\delta$  is a constant shift used when the ratio of the probabilities of  $\mathcal{D}$  and  $\mathcal{D}'$  under the mechanism cannot be bounded by  $e^\epsilon$  (see Definition 2 below). With an arbitrarily given  $\delta$ , a mechanism with a smaller  $\epsilon$  makes *adjacent* databases, ( $\mathcal{D}$  and  $\mathcal{D}'$ ), less distinguishable and hence more private.

**Definition 1 (Adjacency [38])** : Let  $\mathcal{X}$  denote the space of all possible datasets. We say that  $\mathcal{D} \in \mathcal{X}$  and  $\mathcal{D}' \in \mathcal{X}$  are adjacent if they differ on a single element.

**Definition 2 ( $(\epsilon, \delta)$ -Differential Privacy [38])** : The random mechanism  $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{R}$  with domain  $\mathcal{X}$  and range  $\mathcal{R}$  is said to provide  $(\epsilon, \delta)$ -differential privacy, if for any two adjacent datasets  $\mathcal{D}, \mathcal{D}' \in \mathcal{X}$  and for all measurable sets  $\mathcal{S} \subseteq \mathcal{R}$ :

$$\Pr(\mathcal{M}(\mathcal{D}) \in \mathcal{S}) \leq e^\epsilon \Pr(\mathcal{M}(\mathcal{D}') \in \mathcal{S}) + \delta. \quad (34)$$

If  $\delta = 0$ ,  $\mathcal{M}$  is said to satisfy  $\epsilon$ -differential privacy. From Definition 2, we have that a mechanism provides DP if its probability distribution satisfies (34) for some  $\epsilon$  and  $\delta$ . Then, if we seek to design the mechanism to guarantee DP, we need to shape its probability distribution. This is usually done by injecting noise into the data we seek to encode. The noise statistics must be designed in terms of the sensitivity of the data to be encoded. Sensitivity refers to the maximum change of the data due to the difference in a single element of the dataset [38].

**Definition 3 (Sensitivity)** : Given adjacent datasets  $\mathcal{D}, \mathcal{D}' \in \mathcal{X}$ , and a query function  $q : \mathcal{X} \rightarrow \mathcal{R}$  (a deterministic function of datasets) where the output space  $\mathcal{R}$  is equipped with a norm denoted  $\|\cdot\|_{\mathcal{R}}$ , the sensitivity of  $q(\cdot)$  is formulated as  $\Delta_{\mathcal{R}}^q = \sup_{\mathcal{D}, \mathcal{D}'} \|q(\mathcal{D}) - q(\mathcal{D}')\|_{\mathcal{R}}$ .

The differential privacy mechanism  $\mathcal{M}$  must be designed to ensure that the DP condition (34) holds. According to Definition 3, the sensitivity of the query to which this mechanism is applied determines the design of its variables.

## 5.2 Immersion-based Coding Differential Privacy Guarantee

We concretely formulate the problem of designing the variables of privacy coding mechanisms  $\pi_1(\cdot)$  and  $\pi_2(\cdot)$  in (26) and (31) that, at each global iteration, guarantee the privacy of local databases. We wish to design matrices  $\Pi_1, \Pi_2$ , and random variables  $R_1^t$  and  $R_2^t$  such that the privacy mechanisms  $\pi_1(\cdot)$  and  $\pi_2(\cdot)$  for distorting  $\tilde{w}_i^t$  and  $w^t$  are  $(\tilde{\epsilon}, \tilde{\delta})$  and  $(\epsilon', \delta')$ -Differentially private.

**Problem 3 (Element-Wise Differential Privacy)** Given a sequence of desired privacy levels  $(\tilde{\epsilon}, \tilde{\delta})$  and  $(\epsilon', \delta')$ , design the variables of privacy mechanisms in

(26) and (31) such that at global iteration  $t$ , each element of vector  $\tilde{w}_i^t$ ,  $\tilde{w}_{i,j}^t$ , and each element of matrix  $w^t$ ,  $w_{j,m}^t$ ,  $j \in \{1, \dots, \tilde{n}\}$  and  $m \in \{1, \dots, p\}$ , are  $(\tilde{\epsilon}, \tilde{\delta})$  and  $(\epsilon', \delta')$ -Differentially private, respectively, for any measurable  $\mathcal{S} \subseteq \mathbb{R}$ , i.e.,

$$\begin{cases} \mathbb{P}(\tilde{w}_{i,j}^t(\mathcal{D}_i) \in \mathcal{S}) \leq e^{\tilde{\epsilon}} \mathbb{P}(\tilde{w}_{i,j}^t(\mathcal{D}'_i) \in \mathcal{S}) + \tilde{\delta}, \\ \mathbb{P}(w_{j,m}^t(\mathcal{D}_i) \in \mathcal{S}) \leq e^{\epsilon'} \mathbb{P}(w_{j,m}^t(\mathcal{D}'_i) \in \mathcal{S}) + \delta', \\ \text{for adjacent } (\mathcal{D}_i, \mathcal{D}'_i). \end{cases} \quad (35)$$

## 5.3 Solution to Problem 3

As standard in DP, we consider two cases for stochastic processes  $R_1^t$  and  $R_2^t$ , Laplace and Gaussian distributions, and prove DP guarantees for both scenarios.

Starting with the Laplace additive noise scenario, let the independent stochastic process  $R_1^t$  and  $R_2^t$  follow multivariate i.i.d. Laplace distributions with means  $E[R_1^t] =: \mu_1 \in \mathbb{R}^{(\tilde{n}-n) \times p}$  and  $E[R_2^t] =: \mu_2 \in \mathbb{R}^{\tilde{n} \times (p-1)}$ , and covariance matrices  $E[(R_1^t - \mu_1)(R_1^t - \mu_1)^\top] =: \sigma_1 I_{(\tilde{n}-n)}$  and  $E[(R_2^t - \mu_2)(R_2^t - \mu_2)^\top] =: \sigma_2 I_{\tilde{n}}$ , for some  $\sigma_1, \sigma_2 > 0$ , i.e.,  $R_1^t \sim \text{Laplace}(\mu_1, \sigma_1 I_{(\tilde{n}-n)})$  and  $R_2^t \sim \text{Laplace}(\mu_2, \sigma_2 I_{\tilde{n}})$ .

We start with the privacy guarantee for  $w_i^t$ . According to Definition 3, given adjacent local databases  $\mathcal{D}_i, \mathcal{D}'_i \in \mathcal{X}_i$ , where  $\mathcal{X}_i$  denotes the space of all user data sets, the sensitivity of  $w_i^t$  defined is as follows:

$$\Delta_1^{w_i} = \sup_{\mathcal{D}_i, \mathcal{D}'_i} \|w_i^t(\mathcal{D}_i) - w_i^t(\mathcal{D}'_i)\|_1. \quad (36)$$

For simplicity, in what follows, we write  $w_i^t(\mathcal{D}_i)$  and  $w_i^t(\mathcal{D}'_i)$  as  $w_i^t$  and  $w_i^{t'}$ . Because  $R_1^t \sim \text{Laplace}(\mu_1, \sigma_1 I)$ , and given the privacy encoding mechanisms  $\tilde{w}_i^t = \Pi_1 w_i^t + b_1^t$ , with  $b_1^t = N_1 R_1^t \Pi_2^R$  in the extended SIFL, each element of  $\tilde{w}_i^t$  also follows a Laplace distribution:

$$\tilde{w}_{i,j}^t \sim \text{Laplace}\left(\Pi_1^j w_i^t + N_1^j \mu_1 \Pi_2^R, \|\Pi_1^j\|_2 \sigma_1 \|\Pi_2^R\|_2\right), \quad (37)$$

where  $\tilde{w}_{i,j}^t$  is the  $j^{\text{th}}$  element of  $\tilde{w}_i^t$ , and  $\Pi_1^j$  and  $N_1^j$  are the  $j^{\text{th}}$  rows of  $\Pi_1$  and  $N_1$ , respectively. It follows that:

$$\begin{aligned} & \mathbb{P}(\tilde{w}_{i,j}^t(\mathcal{D}_i) \in \mathcal{S}) \\ &= \left( \frac{1}{2 \|\Pi_1^j\|_2 \sigma_1 \|\Pi_2^R\|_2} \right) \int_{\mathcal{S}} e^{-\frac{\|p - (\Pi_1^j w_i^t + N_1^j \mu_1 \Pi_2^R)\|_1}{\|\Pi_1^j\|_2 \sigma_1 \|\Pi_2^R\|_2}} dp \\ &\stackrel{(a)}{\leq} \frac{\|\Pi_1^j(w_i^t - w_i^{t'})\|_1}{2 \|\Pi_1^j\|_2 \sigma_1 \|\Pi_2^R\|_2} \int_{\mathcal{S}} e^{-\frac{\|p - (\Pi_1^j w_i^{t'} + N_1^j \mu_1 \Pi_2^R)\|_1}{\|\Pi_1^j\|_2 \sigma_1 \|\Pi_2^R\|_2}} dp \\ &= e^{\frac{\|\Pi_1^j(w_i^t - w_i^{t'})\|_1}{\|\Pi_1^j\|_2 \sigma_1 \|\Pi_2^R\|_2}} \mathbb{P}(\tilde{w}_{i,j}^t(\mathcal{D}'_i) \in \mathcal{S}), \end{aligned} \quad (38)$$

where inequality (a) follows from the triangle inequality –  $\left\|p - (\Pi_1^j w_i^t + N_1^j \mu_1 \Pi_2^R)\right\|_1 \leq \left\|\Pi_1^j (w_i^t - w_i^{t'})\right\|_1 - \left\|p - (\Pi_1^j w_i^{t'} + N_1^j \mu_1 \Pi_2^R)\right\|_1$ .

Due to the sensitivity relation (36), we have

$$\left\|\Pi_1^j (w_i^t - w_i^{t'})\right\|_1 \leq \left\|\Pi_1^j\right\|_1 \Delta_1^{w_i}. \quad (39)$$

Hence, substituting (39) in (38) implies

$$\mathbb{P}(\tilde{w}_{i,j}^t(\mathcal{D}_i) \in \mathcal{S}) \leq e^{\frac{\|\Pi_1^j\|_1 \Delta_1^{w_i}}{\|N_1^j\|_2 \sigma_1 + \|\Pi_2^R\|_2}} \mathbb{P}(\tilde{w}_{i,j}^t(\mathcal{D}'_i) \in \mathcal{S}). \quad (40)$$

Therefore,  $\epsilon^{\tilde{w}}$ -differential privacy (with  $\delta^{\tilde{w}} = 0$ ) of each local model parameter  $\tilde{w}_{i,j}^t$  for all  $j \in \{1, 2, \dots, \tilde{n}\}$ , is guaranteed for  $\Pi_1$ ,  $N_1$ ,  $\Pi_2^R$ , and  $\sigma_1$  satisfying:

$$\frac{\left\|\Pi_1^j\right\|_1 \Delta_1^{w_i}}{\|N_1^j\|_2 \sigma_1 + \|\Pi_2^R\|_2} \leq \epsilon^{\tilde{w}}. \quad (41)$$

Following the same steps, it can be shown that each element of the encoded global model  $w^t = \Pi_1 w^t \Pi_2 + B_1^t + \Pi_1 \Pi_1^L B_2^t$ ,  $w_{j,m}^t$ , with  $B_1^t = N_1 R_1^t$  and  $B_2^t = R_2^t N_2$  is  $\epsilon'$ -Differentially private in the sense of (35) for  $\Pi_1$ ,  $\Pi_2$ ,  $N_1$ ,  $N_2$ ,  $\sigma_1$ , and  $\sigma_2$  satisfying:

$$\frac{\left\|\Pi_1^j\right\|_1 \Delta_1^w \|\Pi_2^m\|_1}{\|N_1^j\|_2 \sigma_1 + \|\Pi_1^j\|_2 \|\Pi_1^L\|_2 \|N_2\|_2 \sigma_2} \leq \epsilon', \quad (42)$$

for all  $j \in \{1, 2, \dots, \tilde{n}\}$  and  $m \in \{1, 2, \dots, p\}$ , where  $\Pi_1^j$  and  $N_1^j$  are the  $j^{\text{th}}$  rows of matrices  $\Pi_1$  and  $N_1$ , respectively,  $\Pi_2^m$  is the  $m^{\text{th}}$  element of vector  $\Pi_2$ ,  $N_2^m$  is  $m^{\text{th}}$  column of  $N_2$ , and  $\Delta_1^w$  is the sensitivity of the global model given by  $\Delta_1^w = \sup_{\mathcal{D}_i, \mathcal{D}'_i} \|w^t(\mathcal{D}_i) - w^t(\mathcal{D}'_i)\|_1$ . Hence, according to the differential privacy conditions, (41) and (42), to improve privacy guarantees by decreasing  $\epsilon^{\tilde{w}}$  and  $\epsilon^{w'}$ , we need to design  $\Pi_1$  and  $\Pi_2$  as small as possible, while designing  $\sigma_1$  and  $\sigma_2$ , the standard deviations of  $R_1^t$  and  $R_2^t$ , and  $\|N_1^i\|_2$  and  $\|N_2^j\|_2$  as large as possible. From (26) and (31), it is obvious that by choosing small  $\Pi_1$  and  $\Pi_2$ , and large  $\sigma_1$  and  $\sigma_2$ ,  $\tilde{w}_i^t$  and  $w^t$  are close to additive random terms  $b_1^t$  and  $B_1^t + \Pi_1 \Pi_1^L B_2^t$ , and practically independent from  $w_i^t$  and  $w^t$ , respectively.

Note that  $\|N_1^i\|_2$ ,  $i \in \{1, \dots, \tilde{n}_y\}$  and  $\|N_2^j\|_2$ ,  $j \in \{1, \dots, p\}$ , must be nonzero, i.e., we need to design  $N_1$  without zero rows and  $N_2$  without zero columns. The latter is not a technical constraint as, for a given  $N_1$  and  $N_2$  with nonzero rows and columns, respectively,  $\Pi_1$  and  $\Pi_2$  can be obtained by solving the equations  $\Pi_1^L N_1 = \mathbf{0}$  and  $N_2 \Pi_2^R = \mathbf{0}$  and computing the right inverse of  $\Pi_1^L$  and left inverse of  $\Pi_2^R$ , respectively.

The conditions on  $\Pi_1$ ,  $N_1$ , and  $\sigma_1$  to guarantee a desired level of privacy are provided in the following theorem.

**Theorem 1 (Differential Privacy through Laplace additive noises)** Consider given Laplace processes  $R_1^t \sim \text{Laplace}(\mu_1, \sigma_1 I_{(\tilde{n}-n)})$  and  $R_2^t \sim \text{Laplace}(\mu_2, \sigma_2 I_{\tilde{n}})$  with standard deviation  $\sigma_1$  and  $\sigma_2$ , respectively, full-rank matrix  $\Pi_1 \in \mathbb{R}^{\tilde{n} \times n}$ ,  $\Pi_2 \in \mathbb{R}^{1 \times p}$ , matrices  $N_1 \in \mathbb{R}^{\tilde{n} \times (\tilde{n}-n)}$  and  $N_2 \in \mathbb{R}^{(p-1) \times p}$  expanding the kernels of  $\Pi_1^L$  and  $\Pi_2^R$ , respectively, that satisfy the following conditions:

$$\begin{cases} \frac{\left\|\Pi_1^j\right\|_1 \Delta_1^{w_i}}{\|N_1^j\|_2 \sigma_1 + \|\Pi_2^R\|_2} \leq \epsilon^{\tilde{w}}, \\ \frac{\left\|\Pi_1^j\right\|_1 \Delta_1^w \|\Pi_2^m\|_1}{\|N_1^j\|_2 \sigma_1 + \|\Pi_1^j\|_2 \|\Pi_1^L\|_2 \|N_2\|_2 \sigma_2} \leq \epsilon', \end{cases} \quad (43)$$

$\tilde{\epsilon}$  and  $\epsilon'$ -differential privacy guarantee are warranted for each element of vector  $\tilde{w}_i^t$ ,  $\tilde{w}_{i,j}^t$ , and each element of matrix  $w^t$ ,  $w_{j,m}^t$ , respectively, for  $j \in \{1, \dots, \tilde{n}\}$  and  $m \in \{1, \dots, p\}$ .

**Proof:** The proof follows from the discussion provided in this section above. ■

The differential privacy guarantees are also provided when the additive noises in the privacy mechanisms (26) and (31) are Gaussian. Let the independent stochastic processes  $R_1^t$  and  $R_2^t$  following multivariate i.i.d. Gaussian distributions with  $E[R_1^t] = \mathbf{0}$  and  $E[R_2^t] = \mathbf{0}$ , and covariance matrices  $E[(R_1^t)(R_1^t)^T] =: \sigma_1 I_{(\tilde{n}-n)}$  and  $E[(R_2^t)(R_2^t)^T] =: \sigma_2 I_{\tilde{n}}$ , for some  $\sigma_1, \sigma_2 > 0$ , i.e.,  $R_1^t \sim \mathcal{N}(\mathbf{0}, \sigma_1 I_{(\tilde{n}-n)})$  and  $R_2^t \sim \mathcal{N}(\mathbf{0}, \sigma_2 I_{\tilde{n}})$ . In this case, the conditions on  $\Pi_1$ ,  $N_1$ ,  $\Pi_2^R$ ,  $\sigma_1$ , and  $\sigma_2$  to guarantee a desired level of privacy are provided in the following theorem.

**Theorem 2 (Differential Privacy through Gaussian additive noises)** Consider Gaussian processes  $R_1^t \sim \mathcal{N}(\mathbf{0}, \sigma_1 I_{(\tilde{n}-n)})$  and  $R_2^t \sim \mathcal{N}(\mathbf{0}, \sigma_2 I_{\tilde{n}})$  with standard deviation of Gaussian noises  $\sigma_1$  and  $\sigma_2$ , full rank matrix  $\Pi_1 \in \mathbb{R}^{\tilde{n} \times n}$ ,  $\Pi_2 \in \mathbb{R}^{1 \times p}$ , matrices  $N_1 \in \mathbb{R}^{\tilde{n} \times (\tilde{n}-n)}$  and  $N_2 \in \mathbb{R}^{(p-1) \times p}$  expanding the kernels of  $\Pi_1^L$  and  $\Pi_2^R$ , respectively, satisfying the following conditions:

$$\begin{cases} (\|N_1^j\|_2 \sigma_1 + \|\Pi_2^R\|_2)^2 - \frac{\|\Pi_1^j\|_2^2 (\Delta_2^{w_i})^2}{2\tilde{\epsilon}} \\ - \frac{\|\Pi_1^j\|_2 \Delta_2^{w_i}}{\tilde{\epsilon}} Q^{-1}(\tilde{\delta}) (\|N_1^j\|_2 \sigma_1 + \|\Pi_2^R\|_2) \geq 0, \\ (\|N_1^j\|_2 \sigma_1 + \|N_2^m\|_2 \sigma_2)^2 - \frac{\|\Pi_1^j\|_2^2 (\Delta_2^w)^2 \|\Pi_2^m\|_2^2}{2\tilde{\epsilon}} \\ - \frac{\|\Pi_1^j\|_2 \Delta_2^w \|\Pi_2^m\|_2}{\tilde{\epsilon}} Q^{-1}(\tilde{\delta}) (\|N_1^j\|_2 \sigma_1 \\ + \|\Pi_1^j\|_2 \|\Pi_1^L\|_2 \|N_2\|_2 \sigma_2) \geq 0, \end{cases} \quad (44)$$

with the  $Q$ -function  $Q(x) := \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-u^2/2} du$ , i.e., the tail distribution of the standard normal distribution.

Then,  $(\tilde{\epsilon}, \tilde{\delta})$  and  $(\epsilon', \delta')$ -Differential Privacy guarantees are warranted for each element of  $\tilde{w}_i^t$ ,  $\tilde{w}_{i,j}^t$ , and each element of  $w^{t'}$ ,  $w_{j,m}^{t'}$ , respectively, for  $j \in \{1, \dots, \tilde{n}\}$  and  $m \in \{1, \dots, p\}$ , at global iteration  $t$  of the FL algorithm.

**Proof:** See Appendix 8.1. ■

The inequality condition in (44) shows the relation between privacy levels,  $(\tilde{\epsilon}, \tilde{\delta})$  and  $(\epsilon', \delta')$ , and privacy mechanism design variables  $(\sigma_1, \sigma_2, \Pi_1, \Pi_2, N_1)$ . To obtain a higher level of privacy, which can be achieved by reducing the amount of  $(\tilde{\epsilon}, \tilde{\delta})$  and  $(\epsilon', \delta')$ , we need to choose smaller  $\Pi_1$  and  $\Pi_2$ , larger  $N_1$ , and larger noise standard deviations  $\sigma_1$  and  $\sigma_2$ .

**Remark 3** The sizes of the additive noises required to guarantee a desired level of DP for local and global FL models are multiples of the sensitivities of local databases for local and global models,  $\Delta_l^{w_i}$  and  $\Delta_l^w$  for  $l = 1, 2$ . These sensitivities can be calculated by  $\Delta_l^{w_i} = \frac{2C}{|\mathcal{D}_i|}$  and  $\Delta_l^w = \frac{2C}{|\mathcal{D}|}$  where  $C$  is a clipping threshold for bounding  $w_i$  [12]. Since in the SIFL method the distortion induced by these noises can be removed by the server and clients, the noises do not need to be small. Therefore, we can choose a large clipping threshold to avoid distorting the FL performance.

**Remark 4** In (41), (42), and Theorem 2, we propose the design conditions for the variables of the privacy mechanisms,  $\Pi_1, \Pi_2, \sigma_1$ , and  $\sigma_2$ , to guarantee  $(\tilde{\epsilon}, \tilde{\delta})$  and  $(\epsilon', \delta')$ -DP for local and global models. It has been shown that a differentially private algorithm is perfectly secret if the set of differential privacy levels is reached zero [28]. Hence, to have perfect secrecy for coding mechanisms in SIFL methods, we need  $(\tilde{\epsilon}, \tilde{\delta}, \epsilon', \delta') \rightarrow 0$ . Although this cannot be achieved due to the numerical limits,  $(\tilde{\epsilon}, \tilde{\delta}, \epsilon', \delta')$  can be arbitrarily small by selecting small  $\Pi_1$  and  $\Pi_2$  and large  $\sigma_1$  and  $\sigma_2$ , respectively.

## 6 Simulation Experiments

### 6.1 Experimental Setup

We examine the experimental results for both SIFL methods, on three neural network models, namely the Multi-Layer Perceptron (MLP) and two different Convolutional Neural Networks (CNN). Our investigation involves utilizing different optimization tools [32], e.g., Adam, SGD, and Momentum, on the MNIST [39] and Fashion-MNIST [40] databases. The experimental details are described as follows:

- **Datasets:** We test our algorithm on the MNIST database for handwritten digit recognition and Fashion-MNIST database for Zalando’s article images classification, both containing 60000 training and 10000 testing instances of  $28 \times 28$  size gray-level images and 10 classes.
- **Models:** The MLP model is a feed-forward deep neural network with ReLU units and a softmax layer of

10 classes (corresponding to the ten digits) with two hidden layers containing 200 hidden units containing 199,210 parameters. The first CNN model consists of two  $3 \times 3$  convolutional layers followed by the  $2 \times 2$  max pooling layer and ReLU activation function. The first layer has 32 channels, while the second has 64 channels. The fully connected layer has 128 units and takes the flattened output of the second convolutional layer as the input. The CNN model contains 1,199,882 parameters. The second CNN model (CNN2) has two  $5 \times 5$  convolution layers (the first with 32 channels, the second with 64, each followed with  $2 \times 2$  max pooling), a fully connected layer with 512 units and ReLU activation, and a final softmax output layer, containing 582,026 weight parameters. Cross-entropy is employed as the loss function in all three networks.

- **Optimization tools:** As optimization algorithms, the SGD, Momentum, and Adam optimizers with learning rates 0.01, 0.01, and 0.001, and local epoch  $K = 2$ ,  $K = 2$ , and  $K = 1$ , respectively, are employed. To be able to compare the effect of the immersion-based coding algorithm, the immersed optimizers based on the immersion coding given in Proposition 1, target SGD, target Momentum, and target Adam optimizers are employed for training SIFL models.

We set the number of clients to  $N_c = 10$ . Our implementation uses Keras with a Tensorflow backend on an HP laptop with A100 GPU and 16 GB RAM.

We implement various FL algorithms through standard FL (FL), the SIFL Method (SIFL M1) given in Algorithm 1, and the extended SIFL Method (SIFL M2) in Algorithm 2. To be able to implement SIFL M1 and SIFL M2, the variables of the encoding mechanisms and the target optimizer are designed by selecting small full-rank matrices  $\Pi_1$  and  $\Pi_2$  with appropriate dimensions. We compute the base  $N_1$  of the kernel of  $\Pi_1^L$  and the base  $N_2$  of the kernel of  $\Pi_2^R$ . The random processes  $R_1^t$  and  $R_2^t$  are defined at every round as multivariate Gaussian variables with large covariances. The immersed dimensions of model parameters are shown in Table 3.

To calculate DP guarantees for local and global models according to Theorem 1, first we determine the sensitivities of local and global models  $\Delta_1^{w_i}$  and  $\Delta_1^w$ , which according to Remark 3 and considering the clipping threshold  $C = 1000$ , number of clients  $N_c = 10$ , and the size of dataset  $|\mathcal{D}| = 60000$  and local datasets  $|\mathcal{D}_i| = 6000$ , can be calculated as  $\Delta_1^{w_i} = \frac{2C}{|\mathcal{D}_i|} = 0.33$  and  $\Delta_1^w = \frac{2C}{|\mathcal{D}|} = 0.033$ . Then, based on Theorem 1, considering encoding matrices  $\Pi_1$  and  $\Pi_2$ , with  $\|\Pi_1^j\|_1 = 10^{-3}$ ,  $\|\Pi_2^R\|_2 = 10^3$ ,  $\|N_1^j\|_2 = 10^3$ ,  $\|\Pi_2^m\|_2 = 10^{-3}$  and Laplace processes  $R_1^t \sim \text{Laplace}(\mathbf{0}, \sigma_1 I)$  and  $R_2^t \sim \text{Laplace}(\mathbf{0}, \sigma_2 I)$  with standard deviations  $\sigma_1 = \sigma_2 = 10^3$ , the  $\tilde{\epsilon}$  and  $\epsilon'$ -DP guarantees with  $\tilde{\epsilon} = 1e - 12$  and  $\epsilon' = 1e - 13$  are warranted for each elements of local and global models, which are very high-levels of DP-guarantee. Note that since in this method, the privacy noises are removed by the server and clients, they do not need to be small.

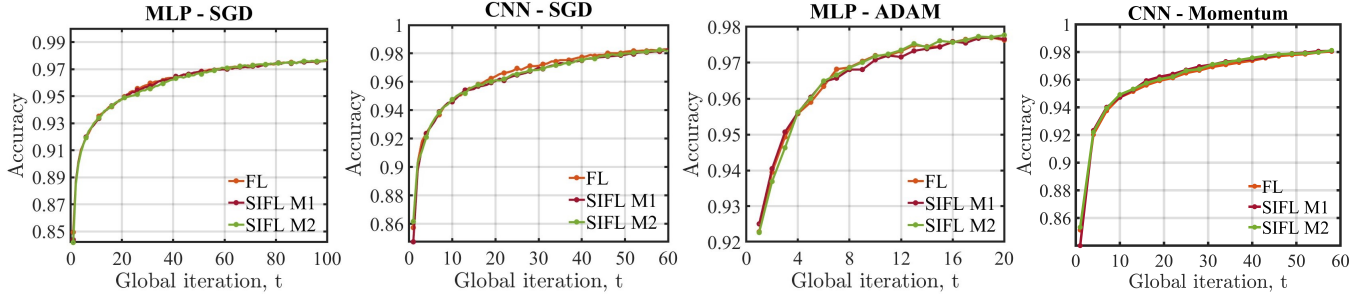


Fig. 2. The comparison of the accuracy of FL network in each iteration with and without the proposed privacy mechanism.

Table 3  
Dimensions of model parameters.

Model	$n$	$\tilde{n}$	$n'$
MLP	199,210	199,411	398,822
CNN	1,199,882	1,200,011	2,400,022
CNN2	582,026	582,539	1,165,078

Therefore, high levels of DP guarantees are achieved without distorting the performance of the FL model.

### 6.2 Performance Evaluation

The comparison of training accuracy of the standard FL algorithm and the proposed SIFL M1 and SIFL M2 are shown in Figure 2. The test accuracies are shown for different models (MLP and CNN), using different optimizers (SGD, Adam, and momentum) and their equivalent target optimizers for SIFL M1 and SIFL M2. The comparison of the accuracy of the FL algorithm for the Fashion-MNIST database and CNN2 model is shown in Figure 3. As can be seen, the SIFL M1 and SIFL M2 accuracy is almost the same as the accuracy with no privacy setting in all scenarios, which shows that SIFL can integrate a cryptographic method in the FL system without sacrificing model accuracy and convergence rate. Therefore, there is no need to make a trade-off between privacy and FL performance.

In Figures 4, we investigate the effect of the encoding and decoding operations in SIFL M1 and SIFL M2 on FL training time. As can be seen, for the MLP model, where the number of model parameters is smaller, the additional training time compared to the training time of the original FL is negligible. However, by increasing the number of model parameters in CNN models, the training time increases. The reason is that by increasing the number of model parameters, the size of multiplicative matrices in privacy mechanisms  $\Pi_1$  and  $\Pi_2$  increases. In addition, the model parameters are flattened at every iteration to become a vector to be able to apply privacy mechanisms, which takes more time.

We compare the accuracy of our proposed scheme with a differential privacy-based FL algorithm in Figure 5, in which the distortion induced by the DP noises is not removed from the model. The federated learning algorithm with differential privacy proposed in [12], namely,

noising before model aggregation FL (NbAFL) is employed to compare the accuracy. It should be noted that to be able to calculate the sensitivity of model parameters in federated learning with differential privacy algorithms, a clipping technique is employed to ensure that  $\|w_i^t\| \leq C$  with clipping threshold  $C$ . In standard FL with DP algorithms, if the clipping threshold  $C$  is too small, clipping destroys the intended gradient direction of parameters, and if it is too large, it forces to add more noise to the parameters because of its effect on the sensitivity. However, in SIFL, since the server and the clients can remove the distortion induced by the privacy noises, they do not need to be small. Therefore, we can choose a large clipping threshold to avoid distorting the FL performance. Hence, in this implementation, the clipping threshold for NbAFL is  $C = 10$ , and for SIFL is  $C = 1000$ . Other parameters in this implementation are  $N_C = 50$  and batch size is 1024, CNN2 model with SGD optimizer and learning rate 0.1 are employed. We measure the model accuracy of NbAFL in a given DP levels  $(\epsilon, \delta)$  for  $\delta = 1e-5$  and  $\epsilon = \{1, 5, 10, 20\}$ . To implement SIFL algorithm, we used Gaussian additive noises  $R_1^t \sim \mathcal{N}(\mathbf{0}, \sigma_1 I)$  and  $R_2^t \sim \mathcal{N}(\mathbf{0}, \sigma_2 I)$  with standard deviations  $\sigma_1 = \sigma_2 = 10^3$ . According to Theorem 2 for calculating the DP guarantees for Gaussian noises, considering  $\|\Pi_1^i\|_2 = 10^{-3}$ ,  $\|\Pi_2^R\|_2 = 10^3$ ,  $\|N_1^j\|_2 = 10^3$ ,  $\sigma_1 = \sigma_2 = 10^3$ , sensitivities of local and global models  $\Delta_1^{w_i} = 0.33$  and  $\Delta_1^w = 0.033$ , the  $(\tilde{\epsilon}, \tilde{\delta})$  and  $(\epsilon', \delta')$ -DP guarantees for each element of local and global models can be achieved with  $\tilde{\epsilon} = 1e-11$ ,  $\tilde{\delta} = 1e-5$ ,  $\epsilon' = 1e-13$ , and  $\delta' = 1e-5$  which are very high levels of DP guaran-

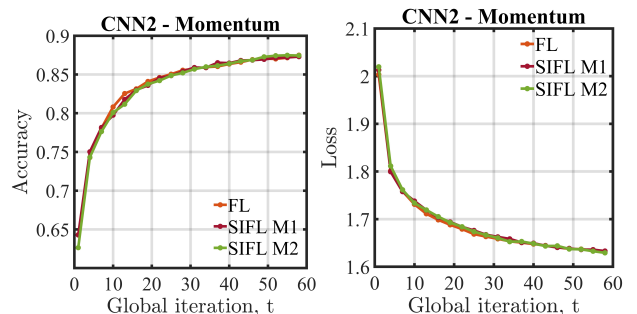


Fig. 3. The comparison of the accuracy and loss of FL with and without privacy for the Fashion-MNIST database.

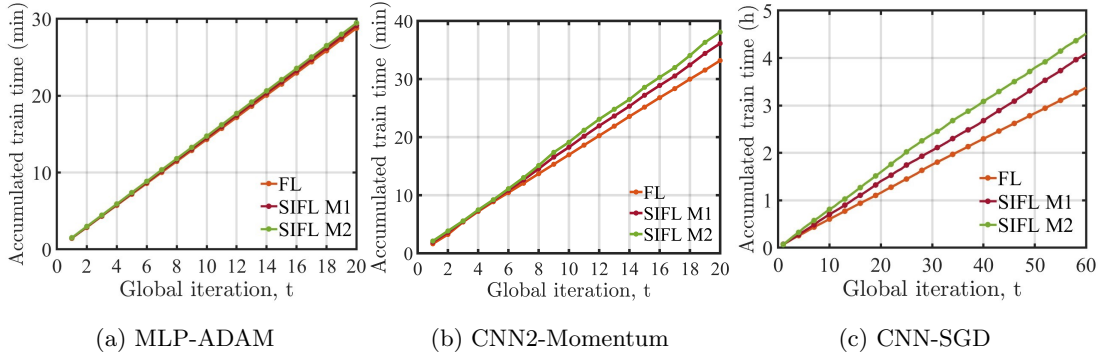


Fig. 4. The comparison of the training time of FL with and without the proposed privacy mechanism.

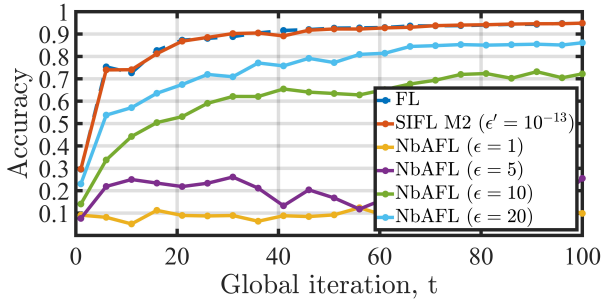


Fig. 5. The comparison of the accuracy of FL, SIFL M2, and NbAFL for various privacy levels  $\epsilon = 1, 5, 10, 20$ .

tee. As can be seen in Figure 5, a higher level of DP guarantee in a standard DP-based FL algorithm like NbAFL would affect the accuracy and convergence rate significantly due to the distorting noises, while the proposed SIFL algorithm can provide a very high level of DP guarantee without losing the model accuracy.

## 7 Conclusions

In this paper, we have developed a System Immersion Federated Learning, SIFL, as a privacy-preserving FL framework built on the synergy of random coding and system immersion tools from control theory to protect privacy of the clients' data in federated learning. The core idea is to treat the Gradient descent optimization algorithm employed in the standard FL process as a dynamical system that we seek to immerse into a higher-dimensional system. We have devised a synthesis procedure to design the dynamics of a coding scheme for privacy and an immersed higher-dimensional optimization algorithm called target optimizer such that model parameters of the standard optimization algorithm are immersed/embedded in its parameters, and it operates on randomly encoded higher-dimensional model parameters. Random coding was formulated at the server as a random change of coordinates that maps the original private parameters of the FL model to a higher-dimensional space. Such coding enforces that the target optimization algorithm converges to an encoded higher-dimensional

version of the model parameters of the original optimization algorithm that can be decoded at the server after model aggregation.

SIFL provides the same accuracy and convergence rate as the standard FL (i.e., when no coding is employed to protect against data inference), reveals no information about clients' data, can be applied to large-scale models, is computationally efficient, and offers any desired level of differential privacy without degrading the FL accuracy and performance. The simulation results of SIFL are presented to illustrate the performance of our tool. These results demonstrate that SIFL provides the same accuracy and convergence rate as the standard FL with a negligible computation cost.

## References

- [1] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial intelligence and statistics, PMLR, 2017, pp. 1273–1282.
- [2] T. Li, A. K. Sahu, A. Talwalkar, V. Smith, Federated learning: Challenges, methods, and future directions, IEEE Signal Processing Magazine 37 (3) (2020) 50–60.
- [3] P. P. Shinde, S. Shah, A review of machine learning and deep learning applications, in: 2018 Fourth international conference on computing communication control and automation (ICCUBEA), IEEE, 2018, pp. 1–6.
- [4] M. Nasr, R. Shokri, A. Houmansadr, Comprehensive privacy analysis of deep learning, in: Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP), 2018, pp. 1–15.
- [5] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: 2017 IEEE symposium on security and privacy (SP), IEEE, 2017, pp. 3–18.
- [6] J. So, R. E. Ali, B. Güler, J. Jiao, A. S. Avestimehr, Securing secure aggregation: Mitigating multi-round privacy leakage in federated learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 9864–9873.
- [7] M. Fredrikson, S. Jha, T. Ristenpart, Model inversion attacks that exploit confidence information and basic countermeasures, in: Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, 2015, pp. 1322–1333.

- [8] Y. Aono, T. Hayashi, L. Wang, S. Moriai, et al., Privacy-preserving deep learning via additively homomorphic encryption, *IEEE Transactions on Information Forensics and Security* 13 (5) (2017) 1333–1345.
- [9] X. Yin, Y. Zhu, J. Hu, A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions, *ACM Computing Surveys (CSUR)* 54 (6) (2021) 1–36.
- [10] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, L. Zhang, Deep learning with differential privacy, in: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.
- [11] R. Shokri, V. Shmatikov, Privacy-preserving deep learning, in: *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1310–1321.
- [12] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, H. V. Poor, Federated learning with differential privacy: Algorithms and performance analysis, *IEEE Transactions on Information Forensics and Security* 15 (2020) 3454–3469.
- [13] C. Liu, K. H. Johansson, Y. Shi, Distributed empirical risk minimization with differential privacy, *Automatica* 162 (2024) 111514.
- [14] Y. Wang, T. Başar, Decentralized nonconvex optimization with guaranteed privacy and accuracy, *Automatica* 150 (2023) 110858.
- [15] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, K. Seth, Practical secure aggregation for privacy-preserving machine learning, in: *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1175–1191.
- [16] G. Xu, H. Li, S. Liu, K. Yang, X. Lin, Verifynet: Secure and verifiable federated learning, *IEEE Transactions on Information Forensics and Security* 15 (2019) 911–926.
- [17] P. Mohassel, Y. Zhang, Secureml: A system for scalable privacy-preserving machine learning, in: *2017 IEEE Symposium on Security and Privacy (SP)*, IEEE, 2017, pp. 19–38.
- [18] V. Mugunthan, A. Polychroniadou, D. Byrd, T. H. Balch, Smpai: Secure multi-party computation for federated learning, in: *Proceedings of the NeurIPS 2019 Workshop on Robust AI in Financial Services*, Vol. 21, MIT Press Cambridge, MA, USA, 2019.
- [19] J. Ma, S.-A. Naas, S. Sigg, X. Lyu, Privacy-preserving federated learning based on multi-key homomorphic encryption, *International Journal of Intelligent Systems* (2022).
- [20] J. Li, X. Kuang, S. Lin, X. Ma, Y. Tang, Privacy preservation for machine learning training and classification based on homomorphic encryption schemes, *Information Sciences* 526 (2020) 166–179.
- [21] M. Asad, A. Moustafa, T. Ito, Fedopt: Towards communication efficiency and privacy preservation in federated learning, *Applied Sciences* 10 (8) (2020) 2864.
- [22] Y. Lu, M. Zhu, Privacy preserving distributed optimization using homomorphic encryption, *Automatica* 96 (2018) 314–325.
- [23] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, R. Zhang, Y. Zhou, A hybrid approach to privacy-preserving federated learning, in: *Proceedings of the 12th ACM workshop on artificial intelligence and security*, 2019, pp. 1–11.
- [24] Y. Wang, H. V. Poor, Decentralized stochastic optimization with inherent privacy protection, *IEEE Transactions on Automatic Control* 68 (4) (2022) 2293–2308.
- [25] R. Xu, N. Baracaldo, Y. Zhou, A. Anwar, H. Ludwig, Hybridalpha: An efficient approach for privacy-preserving federated learning, in: *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, 2019, pp. 13–23.
- [26] C. A. Choquette-Choo, N. Dullerud, A. Dziedzic, Y. Zhang, S. Jha, N. Papernot, X. Wang, Capc learning: Confidential and private collaborative learning, *arXiv preprint arXiv:2102.05188* (2021).
- [27] A. Astolfi, R. Ortega, Immersion and invariance: A new tool for stabilization and adaptive control of nonlinear systems, *IEEE Transactions on Automatic control* 48 (4) (2003) 590–606.
- [28] H. Hayati, N. van de Wouw, C. Murguia, Privacy in cloud computing through immersion-based coding, *arXiv preprint arXiv:2403.04485* (2024).
- [29] H. Hayati, C. Murguia, N. van de Wouw, Privacy-preserving federated learning via system immersion and random matrix encryption, in: *2022 IEEE 61st Conference on Decision and Control (CDC)*, IEEE, 2022, pp. 6776–6781.
- [30] H. Hayati, S. Heijmans, L. Persoon, C. Murguia, N. van de Wouw, Mo-0304 privacy-preserving federated learning for radiotherapy applications, *Radiotherapy and Oncology* 182 (2023) S238–S240.
- [31] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, D. Bacon, Federated learning: Strategies for improving communication efficiency, *arXiv preprint arXiv:1610.05492* (2016).
- [32] S. Ruder, An overview of gradient descent optimization algorithms, *arXiv preprint arXiv:1609.04747* (2016).
- [33] A. Astolfi, D. Karagiannis, R. Ortega, *Nonlinear and adaptive control with applications*, Vol. 187, Springer, 2008.
- [34] A. Isidori, C. Byrnes, Output regulation of nonlinear systems, *IEEE Transactions on Automatic control* 35 (2) (1990) 131–140.
- [35] F. Delli Priscoli, C. Byrnes, A. Isidori, *Output regulation of uncertain nonlinear systems* (1997).
- [36] A. Haro, M. Canadell, J.-L. Figueras, A. Luque, J.-M. Mondelo, *The parameterization method for invariant manifolds*, *Applied mathematical sciences* 195 (2016).
- [37] M. W. Hirsch, S. Smale, R. L. Devaney, *Differential equations, dynamical systems, and an introduction to chaos*, Academic press, 2012.
- [38] C. Dwork, Differential privacy: A survey of results, in: *Theory and Applications of Models of Computation*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 1–19.
- [39] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *Proceedings of the IEEE* 86 (11) (1998) 2278–2324.
- [40] H. Xiao, K. Rasul, R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, *arXiv preprint arXiv:1708.07747* (2017).

## 8 Appendix

### 8.1 Proof of Theorem 2

We start with the privacy guarantee for  $w_i^t$ . According to Definition 3, given the adjacent local databases  $\mathcal{D}_i, \mathcal{D}'_i \in$



$\mathcal{X}_i$ , the sensitivity of  $w_i^t$  can be defined as follows:

$$\Delta_2^{w_i} = \sup_{\mathcal{D}_i, \mathcal{D}'_i} \|w_i^t(\mathcal{D}_i) - w_i^t(\mathcal{D}'_i)\|_2. \quad (45)$$

For simplicity, we show  $w_i^t(\mathcal{D}_i)$  and  $w_i^t(\mathcal{D}'_i)$  by  $w_i^t$  and  $w_i^{t'}$ . Because  $R_1^t \sim \mathcal{N}(\mathbf{0}, \sigma_1 I)$ , and given the privacy encoding mechanisms  $\tilde{w}_i^t = \Pi_1 w_i^t + b_1^t$ , with  $b_1^t = N_1 R_1^t \Pi_2^R$  in the extended SIFL method, each element of  $\tilde{w}_i^t$  follows a Gaussian distribution as  $\tilde{w}_{i,j}^t \sim \mathcal{N}(\Pi_1^j w_i^t, \|N_1^j\|_2 \sigma_1 \|\Pi_2^R\|_2)$ , where  $\Pi_1^j$  and  $N_1^j$  are the  $j$ -th rows of  $\Pi_1$  and  $N_1$ , respectively. We have

$$\begin{aligned} \mathbb{P}(\tilde{w}_{i,j}^t(\mathcal{D}_i)) &= \frac{1}{(2\pi\bar{\sigma}^2)^{\frac{1}{2}}} \int_{\mathcal{S}} e^{-\frac{\|p - \Pi_1^j w_i^t\|_2^2}{2(\|N_1^j\|_2 \sigma_1 \|\Pi_2^R\|_2)^2}} dp \\ &= \frac{1}{(2\pi\bar{\sigma}^2)^{\frac{1}{2}}} \int_{\mathcal{S}} e^{-\frac{\|p - \Pi_1^j w_i^t - \Pi_1^j v\|_2^2}{2(\|N_1^j\|_2 \sigma_1 \|\Pi_2^R\|_2)^2}} dp \\ &= \frac{1}{(2\pi\bar{\sigma}^2)^{\frac{1}{2}}} \int_{\mathcal{S}} e^{-\frac{\|p - \Pi_1^j w_i^{t'}\|_2^2}{2(\|N_1^j\|_2 \sigma_1 \|\Pi_2^R\|_2)^2}} e^{\frac{2(p - \Pi_1^j w_i^{t'}) \Pi_1^j v - \|\Pi_1^j v\|_2^2}{2(\|N_1^j\|_2 \sigma_1 \|\Pi_2^R\|_2)^2}} dp \\ &= \frac{1}{(2\pi\bar{\sigma}^2)^{\frac{1}{2}}} \int_{\mathcal{S} \cap A_\epsilon} e^{-\frac{\|p - \Pi_1^j w_i^{t'}\|_2^2}{2(\|N_1^j\|_2 \sigma_1 \|\Pi_2^R\|_2)^2}} e^{\frac{2(p - \Pi_1^j w_i^{t'}) \Pi_1^j v - \|\Pi_1^j v\|_2^2}{2(\|N_1^j\|_2 \sigma_1 \|\Pi_2^R\|_2)^2}} dp \\ &\quad + \frac{1}{(2\pi\bar{\sigma}^2)^{\frac{1}{2}}} \int_{\mathcal{S} \cap A_\epsilon^c} e^{-\frac{\|p - \Pi_1^j w_i^{t'}\|_2^2}{2(\|N_1^j\|_2 \sigma_1 \|\Pi_2^R\|_2)^2}} dp, \end{aligned} \quad (46)$$

where  $v := w_i^t - w_i^{t'}$ ,  $\bar{\sigma} = (\|N_1^j\|_2 \sigma_1 \|\Pi_2^R\|_2)$ ,

$A_\epsilon = \{p \in \mathbb{R} : \frac{2(p - w_i^{t'}) \Pi_1^j v - \|\Pi_1^j v\|_2^2}{2(\|N_1^j\|_2 \sigma_1 \|\Pi_2^R\|_2)^2} \leq \tilde{\epsilon}\}$  and  $A_\epsilon^c$  denotes its complement. By the definition of  $A_\epsilon$ , the first term of the last expression is bounded by

$$\frac{e^{\tilde{\epsilon}}}{(2\pi\bar{\sigma}^2)^{\frac{1}{2}}} \int_{\mathcal{S}} e^{-\frac{\|p - w_i^{t'}\|_2^2}{2(\|N_1^j\|_2 \sigma_1 \|\Pi_2^R\|_2)^2}} dp = e^{\tilde{\epsilon}} \mathbb{P}(\tilde{w}_{i,j}^t(\mathcal{D}'_i) \in \mathcal{S}). \quad (47)$$

The second integral term is bounded by

$$\frac{1}{(2\pi\bar{\sigma}^2)^{\frac{1}{2}}} \int_{\mathbb{R}} e^{-\frac{\|p - \Pi_1^j w_i^t\|_2^2}{2(\|N_1^j\|_2 \sigma_1 \|\Pi_2^R\|_2)^2}} \mathbf{1}_{\{2(p - \Pi_1^j w_i^{t'}) \Pi_1^j v \geq \|\Pi_1^j v\|_2^2 + 2\tilde{\epsilon}\bar{\sigma}^2\}} dp, \quad (48)$$

which, after the change of variable  $y = (p - \Pi_1^j w_i^t)/\bar{\sigma}$ , can be rewritten

$$\begin{aligned} &\frac{1}{(2\pi)^{\frac{1}{2}}} \int_{\mathbb{R}} e^{-\frac{\|y\|_2^2}{2}} \mathbf{1}_{\{2(\bar{\sigma}y + \Pi_1^j v) \Pi_1^j v \geq \|\Pi_1^j v\|_2^2 + 2\tilde{\epsilon}\bar{\sigma}^2\}} dy \\ &= \frac{1}{(2\pi)^{\frac{1}{2}}} \int_{\mathbb{R}} e^{-\frac{\|y\|_2^2}{2}} \mathbf{1}_{\left\{y \geq \frac{\tilde{\epsilon}\bar{\sigma}}{\|\Pi_1^j v\|_2} - \frac{\|\Pi_1^j v\|_2}{2\bar{\sigma}}\right\}} dy. \end{aligned} \quad (49)$$

This last expression is  $\mathbb{P}\left(Y \geq \frac{\tilde{\epsilon}\bar{\sigma}}{\|\Pi_1^j v\|_2} - \frac{\|\Pi_1^j v\|_2}{2\bar{\sigma}}\right) \leq \tilde{\delta}$ , for  $Y \sim \mathcal{N}(0, 1)$ . We are then led to set  $\bar{\sigma}$  sufficiently

large so that  $\mathbb{P}\left(Y \geq \tilde{\epsilon}\bar{\sigma}/\|\Pi_1^j v\|_2 - \|\Pi_1^j v\|_2/2\bar{\sigma}\right) \leq \tilde{\delta}$ , that is,  $Q\left(\tilde{\epsilon}\bar{\sigma}/\|\Pi_1^j v\|_2 - \|\Pi_1^j v\|_2/2\bar{\sigma}\right) \leq \tilde{\delta}$ . Because  $Q^{-1}$  is monotonically decreasing, we have the condition  $\frac{\tilde{\epsilon}\bar{\sigma}}{\|\Pi_1^j v\|_2} - \frac{\|\Pi_1^j v\|_2}{2\bar{\sigma}} \geq Q^{-1}(\tilde{\delta})$ , which is equivalent to

$$\bar{\sigma}^2 - \bar{\sigma} \frac{\|\Pi_1^j v\|_2}{\tilde{\epsilon}} Q^{-1}(\tilde{\delta}) - \frac{\|\Pi_1^j v\|_2^2}{2\tilde{\epsilon}} \geq 0. \quad (50)$$

From Definition 3, (50) can be converted to:

$$\bar{\sigma}^2 - \bar{\sigma} \frac{\|\Pi_1^j\|_2 \Delta_2^w}{\tilde{\epsilon}} Q^{-1}(\tilde{\delta}) - \frac{\|\Pi_1^j\|_2^2 (\Delta_2^w)^2}{2\tilde{\epsilon}} \geq 0. \quad (51)$$

Substituting  $\bar{\sigma} = (\|N_1^j\|_2 \sigma_1 \|\Pi_2^R\|_2)$  into (51), to have  $(\tilde{\epsilon}, \tilde{\delta})$ -DP guarantee for local models,  $\Pi_1$ ,  $\Pi_2$ ,  $N_1$ , and  $\sigma_1$  should be designed to satisfy the following inequality:

$$\begin{aligned} &(\|N_1^j\|_2 \sigma_1 \|\Pi_2^R\|_2)^2 - \frac{\|\Pi_1^j\|_2 \Delta_2^w}{\tilde{\epsilon}} Q^{-1}(\tilde{\delta}) (\|N_1^j\|_2 \sigma_1 \|\Pi_2^R\|_2) \\ &\quad - \frac{\|\Pi_1^j\|_2^2 (\Delta_2^w)^2}{2\tilde{\epsilon}} \geq 0. \end{aligned} \quad (52)$$

Following the same reasoning, we define the sensitivity of the global model as  $\Delta_2^w = \sup_{\mathcal{D}_i, \mathcal{D}'_i} \|w^t(\mathcal{D}_i) - w^t(\mathcal{D}'_i)\|_2$ . For the extended SIFL method, it can be shown that each element of the encoded global model  $w^{t'} = \Pi_1 w^t \Pi_2 + B_1^t + \Pi_1 \Pi_1^L B_2^t$ ,  $w_{j,m}^{t'}$ , with  $B_1^t = N_1 R_1^t$  and  $B_2^t = R_2^t N_2$  is  $(\epsilon', \delta')$ -Differentially private in (35) for:

$$\begin{aligned} &(\|N_1^j\|_2 \sigma_1 + \|N_2^m\|_2 \sigma_2)^2 - \frac{\|\Pi_1^j\|_2^2 (\Delta_2^w)^2 \|\Pi_2^m\|_2^2}{2\epsilon'} \\ &\quad - \frac{\|\Pi_1^j\|_2 \Delta_2^w \|\Pi_2^m\|_2}{\epsilon'} Q^{-1}(\delta') (\|N_1^j\|_2 \sigma_1) \\ &\quad + \|\Pi_1^j\|_2 \|\Pi_1^L\|_2 \|N_2\|_2 \sigma_2 \geq 0, \end{aligned} \quad (53)$$

for all  $j \in \{1, 2, \dots, \tilde{n}\}$  and  $m \in \{1, 2, \dots, p\}$ , where  $N_2^m$  and  $\Pi_2^m$  are the  $m^{\text{th}}$  column of  $N_2$  and  $m^{\text{th}}$  element of vector  $\Pi_2$ , respectively.  $\blacksquare$