Advancing Personalized Federated Learning: Group Privacy, Fairness, and Beyond

Filippo Galli, Kangsoo Jung, Sayan Biswas, Catuscia Palamidessi, Tommaso Cucinotta

SN Computer Science (2023) 4:831
https://doi.org/10.1007/s42979-023-02292-0

ORIGINAL RESEARCH

Advancing Personalized Federated Learning: Group Privacy, Fairness, and Beyond

Filippo Galli[1,3] · Kangsoo Jung[2] · Sayan Biswas[2,4] · Catuscia Palamidessi[2,4] · Tommaso Cucinotta[3]

Received: 30 June 2023 / Accepted: 31 August 2023

Abstract

Federated learning (FL) is a framework for training machine learning models in a distributed and collaborative manner. During training, a set of participating clients process their data stored locally, sharing only updates of the statistical model's parameters obtained by minimizing a cost function over their local inputs. FL was proposed as a stepping-stone towards privacy-preserving machine learning, but it has been shown to expose clients to issues such as leakage of private information, lack of personalization of the model, and the possibility of having a trained model that is fairer to some groups of clients than to others. In this paper, the focus is on addressing the triadic interaction among personalization, privacy guarantees, and fairness attained by trained models within the FL framework. Differential privacy and its variants have been studied and applied as cutting-edge standards for providing formal privacy guarantees. However, clients in FL often hold very diverse datasets representing heterogeneous communities, making it important to protect their sensitive and personal information while still ensuring that the trained model upholds the aspect of fairness for the users. To attain this objective, a method is put forth that introduces group privacy assurances through the utilization of d-privacy (aka metric privacy). d-privacy represents a localized form of differential privacy that relies on a metric-oriented obfuscation approach to maintain the original data's topological distribution. This method, besides enabling personalized model training in a federated approach and providing formal privacy guarantees, possesses significantly better group fairness measured under a variety of standard metrics than a global model trained within a classical FL template. Theoretical justifications for the applicability are provided, as well as experimental validation on real-world datasets to illustrate the working of the proposed method.

Keywords Federated learning · Metric privacy · Personalized models · Fairness

This article is part of the topical collection "Recent Trends on Information Systems Security and Privacy" guest edited by Steven Furnell and Paolo Mori.

✉ Filippo Galli
filippo.galli@sns.it

✉ Kangsoo Jung
gangsoo.zeong@inria.fr

✉ Sayan Biswas
sayan.biswas@inria.fr; bizwas05@gmail.com

✉ Catuscia Palamidessi
catuscia@lix.polytechnique.fr

✉ Tommaso Cucinotta
tommaso.cucinotta@santannapisa.it

[1] Scuola Normale Superiore, Pisa, Italy

[2] INRIA, Palaiseau, France

[3] Scuola Superiore Sant'Anna, Pisa, Italy

[4] École Polytechnique, Palaiseau, France

Introduction

The widespread collection of user data in modern machine learning has raised concerns regarding privacy violations and the potential disclosure of sensitive personal information [1, 2]. To address these concerns, Federated Learning [3] was introduced as a collaborative machine learning paradigm, where users' devices train a global predictive model without transmitting raw data to a central server. While FL offers promises of preserving user privacy and maintaining model performance, the heterogeneity of data distributions among clients can lead to challenges such as reduced model utility and convergence issues during training. In response, personalized federated learning approaches have emerged, aiming to tailor models to clusters of users with similar data distributions [4–6].

Furthermore, it has been demonstrated that avoiding the release of users' raw data alone does not provide sufficient

Published online: 28 October 2023

protection against potential privacy violations [7–9]. To address this issue, researchers have explored the application of differential privacy (DP) [10, 11] to federated learning, providing privacy guarantees for users participating in the optimization process. DP mechanisms introduce randomness in the model updates released by clients, making each user's contribution to the final model probabilistically indistinguishable up to a certain likelihood factor. To bound this factor, the domain of secrets (i.e., the parameter space in FL) is artificially constrained, either to offer central [12, 13] or local DP guarantees [14, 15]. However, constraining the optimization process to a subset of $\mathbb{R}^n$ can have negative effects, such as when the optimal model parameters for a particular cluster of users lie outside such a bounded domain.

To address the challenges of personalization and local privacy protection, this work proposes the adoption of a more general notion of DP called d-privacy or metric-based privacy [16] which has been in the spotlight of late mainly in the context of location-privacy [17–19]. This concept of privacy does not require a bounded domain and provides guarantees based on the distance between any two points in the parameter space. Therefore, assuming that clients with similar data distributions have similar optimal fitting parameters, d-privacy offers strong indistinguishability guarantees. Conversely, privacy guarantees degrade gracefully for clients with significantly different data distributions.

In addition to addressing privacy concerns in personalised FL as was studied in [20], this work extends the analysis and investigates the impact of the proposed method on fairness aspects in federated model training. As machine learning-based decision systems become more prevalent, it has become apparent that many of these systems exhibit gender and racial biases that disproportionately affect minority populations [21, 22]. Therefore, beyond protecting user privacy, it is crucial to explore cutting-edge machine learning algorithms that can potentially mitigate this pervasive lack of fairness among participating clients. However, systems aiming to protect privacy while ensuring fairness often involve a trade-off between the two [23]. This trade-off arises because privacy protection techniques based on DP tend to minimize the impact of outliers or minorities within the overall dataset. In other words, the application of d-privacy, a metric-based generalization of DP, to personalized FL could potentially compromise the fairness of the machine learning model. Building upon [20], this paper presents extensive experimental results demonstrating that the use of personalized FL under group privacy guarantees not only significantly improves fairness compared to the classical (non-personalized) FL framework, but it also maintains a relatively small trade-off between privacy and fairness.

In summary, the contributions of this paper are the following: it extends the work pursued in [20] (points 1 and 2) and it investigates the implications of our proposal on the fairness of the model (point 3):

1. A novel algorithm is put forward for collaborative training of machine learning models, leveraging advanced techniques for model personalization and addressing user privacy concerns by formalizing privacy guarantees in terms of d-privacy.
2. This research focuses on studying the Laplace mechanism under Euclidean distance, and providing a closed-form expression for its generalization in $\mathbb{R}^n$, as well as an efficient sampling procedure.
3. It shows that personalized federated learning under formal privacy guarantees improves group fairness significantly compared to the non-personalized federated learning framework and, hence, establishes that this method enhances the trade-off between privacy and fairness.

The rest of this paper is organized as follows. Section "Background" introduces the relevant foundations of federated learning, differential privacy, and fairness notions. Section "Related Works" discusses the related works for our research. Section "An Algorithm for Private and Personalized Federated Learning" explains the proposed algorithm for personalized federated learning with group privacy. Section "Experiments" illustrates how the proposed method works in terms of privacy and fairness, and Section "Conclusion" provides our concluding remarks.

Background

Personalized Federated Learning

The problem of personalized federated learning falls within the framework of stochastic optimization, and the notation from [4] is adopted here to determine the set of minimizers $\theta_j^* \in \mathbb{R}^n$ with $j \in \{1, \dots, k\}$ of the cost functions

$$F(\theta_j) = \mathbb{E}_{z \sim \mathcal{D}_j}\left[f(\theta_j; z)\right], \tag{1}$$

where $\{\mathcal{D}_1, \dots, \mathcal{D}_k\}$ are the data distributions which cannot be accessed directly but only through a collection of client datasets $Z_c = \left\{z | z \sim \mathcal{D}_j, z \in \mathbb{D}\right\}$ for some $j \in \{1, \dots, k\}$ with $c \in C = \{1, \dots, N\}$ the set of clients, and $\mathbb{D}$ a generic domain of data points. C is partitioned in k disjoint sets

$$S_j^* = \{c \in C \mid \forall z \in Z_c,\ z \sim \mathcal{D}_j\} \quad \forall j \in \{1, \ldots, k\} \tag{2}$$

The mapping $c \to j$ is unknown and it is necessary to rely on estimates S_j of the membership of Z_c to compute the empirical cost functions

$$\tilde{F}(\theta_j) = \frac{1}{|S_j|} \sum_{c \in S_j} \tilde{F}_c(\theta_j; Z_c); \quad \tilde{F}_c(\theta_j;\ Z_c) = \frac{1}{|Z_c|} \sum_{z_i \in Z_c} f(\theta;\ z_i) \tag{3}$$

The cost function $f : \mathbb{R}^n \times \mathbb{D} \mapsto \mathbb{R}_{\geq 0}$ is applied on $z \in \mathbb{D}$, parametrized by the vector $\theta_j \in \mathbb{R}^n$. Thus, the optimization aims to find, $\forall j \in \{1, \ldots, k\}$,

$$\tilde{\theta}_j^* = \arg\min_{\theta_j} \tilde{F}(\theta_j) \tag{4}$$

Privacy

d-privacy, introduced in [16], extends the concept of differential privacy (DP) to any domain $\mathcal{X}$, which represents the original data space and is equipped with a distance measure $d : \mathcal{X}^2 \mapsto \mathbb{R}_{\geq 0}$, along with a space of secrets $\mathcal{Y}$. A random mechanism $\mathcal{R} : \mathcal{X} \mapsto \mathcal{Y}$ is considered ϵ-d-private if, for any $x_1, x_2 \in \mathcal{X}$ and measurable $S \subseteq \mathcal{Y}$, the inequality in Eq. (5) holds:

$$\mathbb{P}\big[\mathcal{R}(x_1) \in S\big] \leq e^{\epsilon d(x_1, x_2)} \mathbb{P}\big[\mathcal{R}(x_2) \in S\big] \tag{5}$$

It is important to note that when $\mathcal{X}$ corresponds to the domain of databases and d represents the distance based on the Hamming graph of their adjacency relation, Equation (5) aligns with the standard definition of DP in [10, 11]. However, in this study, $\theta \in \mathbb{R}^n$ is considered as both the domain $\mathcal{X}$ and the space of secrets $\mathcal{Y}$. The primary motivation behind employing d-privacy is to preserve the topology of the parameter distributions among clients. Specifically, it aims to ensure that clients with similar model parameters in the non-privatized space $\mathcal{X}$ will communicate approximate model parameters in the privatized space $\mathcal{Y}$, on average.

Fairness

With the recent surge of interest in building ethical ways to train machine learning models, the topic of fairness in machine learning has been in the spotlight and, correspondingly, various metrics and algorithms to quantify and establish fairness in model training have been studied from a variety of perspectives and in different contexts [24–26]. Most fairness metrics consider the simple case of having a *privileged* group and an *unprivileged* group in the population. Under this assumption, typically one attribute of the dataset is selected as a sensitive attribute (e.g., gender, race, etc.) that defines the privileged and the unprivileged groups. The goal of fairness in machine learning is to ensure fair and non-discriminated results regardless of the membership in a sensitive attribute. The two main notions of fairness considered by the community are individual fairness and group fairness: *Individual fairness* [27] claims that similar individuals should be treated similarly, and *group fairness* requires that different demographic subgroups should receive equal treatment with respect to their sensitive attributes. While both notions of fairness are important, this work focus on group fairness because our goal is to analyze and mitigate the potential bias against certain groups (e.g. demographic groups) through personalization techniques. The following metrics are considered for evaluating group fairness as a part of this work.

In the rest of the paper, $\hat{Y} = 1$, $\hat{Y} = 0$ is used to represent the positive and negative prediction respectively, and $S = 1$, $S = 0$ to represent the privileged and unprivileged group.

The simplest notion of fairness to be proposed was *demographic parity* [27].

Definition 2.1 *Demographic parity* is achieved by a system when the prediction $\hat{Y}$ of the target label Y is statistically independent of the sensitive attributes S, i.e.,

$$\mathbb{P}\big[\hat{Y} = 1 | S = 1\big] = \mathbb{P}\big[\hat{Y} = 1 | S = 0\big] \tag{6}$$

Imposing demographic parity has often a strong negative impact on accuracy, and, consequently, more refined notions were proposed afterwards. In particular, *equalized odds* and *equal opportunity* [28].

Definition 2.2 A system satisfies *equalized odds* if its prediction $\hat{Y}$ is conditionally independent of the sensitive attribute S given the target label Y,

$$\mathbb{P}\big[\hat{Y} = 1 | Y = y, S = 1\big] = \mathbb{P}\big[\hat{Y} = 1 | Y = y, S = 0\big], \quad y \in \{0, 1\} \tag{7}$$

In other words, the notion of equalized odds requires the privileged and unprivileged groups to have equal true positive rates and equal false positive rates.

Equal opportunity is a relaxation of equalized odds, in the sense that it only requires equal true positive rates across the groups.

Definition 2.3 *Equal opportunity* is satisfied by a system if its prediction $\hat{Y}$ is conditionally independent of the sensitive attribute S given the target label Y

$$\mathbb{P}\big[\hat{Y} = 1 | Y = 1, S = 1\big] = \mathbb{P}\big[\hat{Y} = 1 | Y = 1, S = 0\big] \tag{8}$$

In practice, however, it is difficult to obtain perfect equality for any of the aforementioned notions. Hence,

typically the aim is to minimize the absolute value of the difference between the privileged and unprivileged groups, rather than requiring this difference to be exactly zero. For instance, the *demographic parity difference* is defined as

$$\left|\mathbb{P}\left[\hat{Y}=1|S=1\right]-\mathbb{P}\left[\hat{Y}=1|S=0\right]\right| \tag{9}$$

and similarly for the *equalized odd difference* and *equal opportunity difference*.

Table 1 Qualitative comparison with the most relevant prior research on the topic

	[13]	[33]	[14]	[20]	This Work
Central privacy	✓	✓	✓	✓	✓
Local privacy	×	✓	✓	✓	✓
Personalization	×	✓	×	✓	✓
Mild Assumptions on Training	✓	×	✓	✓	✓
Fairness analysis	×	×	×	×	✓

Related Works

Federated optimization has demonstrated suboptimal performance when the local datasets consist of samples from non-congruent distributions, resulting in the inability to simultaneously minimize both client-level and global objectives. In previous studies [4–6], researchers examined various meta-algorithms for personalization, but the assertion of preserving user privacy relies solely on clients releasing updated models or model updates, rather than transferring raw data to the server, which can have significant consequences. To address this issue, several works have focused on the privatization of the (federated) optimization algorithm within the framework of DP [12, 13, 29, 30], which adopt DP to provide defences against an *honest-but-curious* adversary. However, even in this setting, there is no guarantee of protection against sample reconstruction from the local datasets using client updates, as highlighted in [9]. Various strategies have been explored to offer local privacy guarantees, either through cryptographic approaches [31] or within the framework of local DP [14, 32, 33]. Specifically, in [33], the authors tackle the problem of personalized and locally differentially private federated learning, but only for the case of simple convex, 1-Lipschitz cost functions of the inputs. It is worth noting that this assumption is unrealistic in the majority of machine learning models and excludes many statistical modelling techniques, particularly neural networks. Finally, some research focused on designing architectures capable of providing private computing environments for remote users [34], often making use of trusted platform modules, secure processors [35], or similar mechanisms [36] improving efficiency by enforcing encryption on network transmissions, rather than memory accesses. For example, the latter work conceptualizes an architecture that could be leveraged to deploy a server that can only reveal the data being processed to clients that instantiated the server. It shall be noted, however, that cryptographic guarantees of security are orthogonal to the privacy notions of differential privacy and its generalizations. To summarize and provide context around this work, Table 1 provides a qualitative evaluation of relevant research and how the contributions presented in this paper fit among them.

Of late, a great deal of attention has been devoted to studying and understanding the aspects of fairness in machine learning [23, 37–42]. Most of the research on fairness focuses on developing techniques to mitigate bias in machine learning algorithms. These techniques can be categorized into three main approaches: pre-processing, in-processing, and post-processing. Pre-processing techniques [43, 44] aim to generate a less biased dataset by modifying the values or adjusting the sampling process. In the case of in-processing techniques [45, 46], the objective function is optimized while taking into account discrimination-aware regularizers. Post-processing techniques [47, 48] involve adjusting the trained model to produce fairer outcomes. However, it is worth noting that the majority of these studies primarily target centralized machine learning models as opposed to FL. Furthermore, there is a lack of research exploring the interplay between accuracy and fairness [40, 41] or privacy and fairness [23, 49]. In particular, to the best of our knowledge, disproportionately fewer works have focused on investigating the relationship between privacy and fairness. [23] formally proved that privacy and fairness can be at odds with each other with non-trivial accuracy. A few recent works on group fairness in FL have emerged [38, 39] but they do not consider the facet of privacy-fairness trade-off.

An Algorithm for Private and Personalized Federated Learning

Algorithm 1 aims to enable personalized federated learning while ensuring local privacy guarantees to preserve group privacy. In this context, locality refers to the sanitization of client information before it is shared with the server, while group privacy pertains to the notion of indistinguishability within a specific neighbourhood of clients, defined based on a particular distance metric. To clarify our terminology, we provide definitions for *neighbourhood* and *group* as follows:

Algorithm 1 An algorithm for personalized federated learning with formal privacyguarantees in local neighbourhoods.

Input: number of clusters k; initial hypotheses $\theta_j^{(0)}, j \in \{1, \ldots, k\}$; number of rounds T; number of users per round U; number of local epochs E; local step size s; user batch size B_s; noise multiplier ν; local dataset Z_c held by user c.
for $t = \{0, 1, \ldots, T-1\}$ **do** ▷ Server-side loop
  $C^{(t)} \leftarrow \text{SampleUserSubset}(U)$
  $\text{BroadcastParameterVectors}(C^{(t)}; \theta_j^{(t)}, j \in \{1, \ldots, k\})$
  for $c \in C^{(t)}$ **do** in parallel ▷ Client-side loop
    $\bar{j} = \arg\min_{j \in \{1,\ldots,k\}} F_c(\theta_j^{(t)}; Z_c)$
    $\theta_{\bar{j},c}^{(t)} \leftarrow \text{LocalUpdate}(\theta_{\bar{j}}^{(t)}; s; E; Z_c)$
    $\hat{\theta}_{\bar{j},c}^{(t)} \leftarrow \text{SanitizeUpdate}(\theta_{\bar{j},c}^{(t)}; \nu)$
  end for
  $\{S_1, \ldots, S_k\} = \text{k-means}(\hat{\theta}_{\bar{j},c}^{(t)}, c \in C^{(t)}; \theta_j^{(t)}, j \in \{1, \ldots, k\})$
  $\theta_j^{(t+1)} \leftarrow \frac{1}{|S_j|} \sum_{c \in S_j} \hat{\theta}_{\bar{j},c}^{(t)}, \quad \forall j \in \{1, \ldots, k\}$
end for

Definition 4.1 For any model parameterized by $\theta_0, \in, \mathbb{R}^n$, the *r-neighbourhood* is defined as the set of points in the parameter space that are within an L_2 distance of r or less from θ_0, i.e., $\theta \in \mathbb{R}^n : \|\theta_0 - \theta\|_2 \leq r$. Clients whose models are parameterized by $\theta \in \mathbb{R}^n$ within the same r-neighbourhood are considered to be part of the same *group* or *cluster*.

Algorithm 1 is inspired by the Iterative Federated Clustering Algorithm (IFCA) proposed in [4] and extends it by incorporating formal privacy guarantees. The key modifications include the introduction of the `SanitizeUpdate` function, as described in Algorithm 2, and the utilization of k-means for server-side clustering of the updated models.

function SANITIZEUPDATE($\theta_{\bar{j}}^{(t)}; \theta_{\bar{j},c}^{(t)}; \nu$)
  $\delta_c^{(t)} = \theta_{\bar{j},c}^{(t)} - \theta_{\bar{j}}^{(t)}$
  $\varepsilon = \frac{n}{\nu \|\delta_c^{(t)}\|}$
  Sample $\rho \sim \mathcal{L}_{0,\varepsilon}(x)$
  $\hat{\theta}_{\bar{j},c}^{(t)} = \theta_{\bar{j},c}^{(t)} + \rho$
  return $\hat{\theta}_{\bar{j},c}^{(t)}$
end function

Algorithm 2 SanitizeUpdate obfuscates a vector θ ∈ Rn, with a Laplacian noisetuned on the radius of a certain neighbourhood and centered in 0.

The Laplace Mechanism Under Euclidean Distance in $\mathbb{R}^n$

The `SanitizeUpdate` function in Algorithm 2 is based on a generalization of the Laplace mechanism to $\mathbb{R}^n$ under the Euclidean distance, which was originally introduced in [50] for geo-indistinguishability in $\mathbb{R}^2$. The decision to utilize the L_2 norm as the distance measure serves two main purposes.

First, clustering is performed on the vector space $\mathbb{R}^n$ of parameters, using the k-means algorithm, which relies on the Euclidean distance. By defining clusters or groups of users based on the proximity of their model parameters using the L_2 norm, the procedure needs a d-privacy mechanism that obscures the reported values within each group while enabling the server to distinguish among users belonging to different clusters.

Second, the use of equidistant noise vectors in the L_2 norm for sanitizing the parameters ensures equiprobability by construction. This property leads to the same bound on the increase of the cost function in first-order approximation, as demonstrated in Proposition 4.2. The Laplace mechanism under Euclidean distance in the general space $\mathbb{R}^n$ is formally defined in Proposition 4.1.

Proposition 4.1 *Let $\mathcal{L}_\varepsilon : \mathbb{R}^n \mapsto \mathbb{R}^n$ be the Laplace mechanism with distribution $\mathcal{L}_{x_0,\varepsilon}(x) = \mathbb{P}\left[\mathcal{L}_\varepsilon(x_0) = x\right] = Ke^{-\varepsilon d(x,x_0)}$ with $d(.)$ being the Euclidean distance. If $\rho \sim \mathcal{L}_{x_0,\varepsilon}(x)$, then:*

1. *$\mathcal{L}_{x_0,\varepsilon}$ is ε-d-private and $K = \frac{\varepsilon^n \Gamma(\frac{n}{2})}{2\pi^{\frac{n}{2}} \Gamma(n)}$*
2. *$\|\rho\|_2 \sim \gamma_{\varepsilon,n}(r) = \frac{\varepsilon^n e^{-\varepsilon r} r^{n-1}}{\Gamma(n)}$*
3. *The i^{th} component of ρ has variance $\sigma_{\rho_i}^2 = \frac{n+1}{\varepsilon^2}$*

where $\Gamma(n)$ is the Gamma function defined for positive reals as $\int_0^\infty t^{n-1} e^{-t}\, dt$ which reduces to the factorial function whenever $n \in \mathbb{N}$.

Proof The proof can be found in Appendix A of [20]. □

Proposition 4.2 *Let $y = f(x, \theta)$ be the fitting function of a machine learning model parameterized by θ, and $(X, Y) = Z$ the dataset over which the RMSE loss function $F(Z, \theta)$ is to be minimized, with $x \in X$ and $y \in Y$. If $\rho \sim \mathcal{L}_{0,\varepsilon}$, the bound on the increase of the cost function does not depend on the direction of ρ, in first-order approximation, and*:

$$\|F(Z, \theta + \rho)\|_2 - \|F(Z, \theta)\|_2 \leq \left\|J_f(X, \theta)\right\|_2 \|\rho\|_2 + o(\left\|J_f(X, \theta) \cdot \rho\right\|_2) \tag{10}$$

Proof The proof can be found in Appendix A of [20]. □

The results in Proposition 4.1 allow to reduce the problem of sampling a point from Laplace to (i) sampling the norm of such point according to the result in Item 4.1 of Proposition 4.1 and then (ii) sample uniformly a unit (directional) vector from the hypersphere in $\mathbb{R}^n$. Much like DP, d-privacy provides a means to compute the total privacy parameters in case of repeated queries, a result known as the Compositionality Theorem for d-privacy.

Theorem 4.1 *Let $\mathcal{K}_i$ be (ε_i)-d-private mechanism for $i \in \{1, 2\}$. Then their independent composition is $(\varepsilon_1 + \varepsilon_2)$ -d-private.*

Proof The proof can be found in Appendix A of [20]. □

A Heuristic for Defining the Neighbourhood of a Client

During the t-th iteration, when a user c invokes the `SanitizeUpdate` procedure in Algorithm 2, it has already received a set of hypotheses, optimized $\theta_{\tilde{j}}^{(t)}$ (the one that fits best its data distribution), and got $\theta_{\tilde{j},c}^{(t)}$. It is reasonable to assume that clients whose datasets are sampled from the same underlying data distribution $\mathcal{D}_{\tilde{j}}$ will perform an update similar to $\delta_c^{(t)}$. Therefore, points which are within the $\delta_c^{(t)}$ -neighbourhood of $\hat{\theta}_{\tilde{j},c}^{(t)}$ are forced to be indistinguishable. To provide this guarantee, the Laplace mechanism is tuned such that the points within the neighbourhood are $\varepsilon\|\delta_c^{(t)}\|_2$ differentially private. By choosing $\varepsilon = n/(\nu\delta_c^{(t)})$, it results in $\varepsilon\|\delta_c^{(t)}\|_2 = n/\nu$, where ν is referred to as the *noise multiplier*. Notably, a larger value of ν corresponds to a stronger privacy guarantee. This is because the norm of the noise vector sampled from the Laplace distribution follows the distribution specified in Proposition 4.1, with an expected value of $\mathbb{E}\left[\gamma_{\varepsilon,n}(r)\right] = n/\varepsilon$.

Experiments

The following Section discusses a number of experimental validations of Algorithm 1 on different tasks and datasets. Detailed experimental settings are discussed in Appendix B of [20], but we provide here an overview of the hardware and software stacks: All the following experiments are run on a local server running Ubuntu 20.04.3 LTS with an AMD EPYC 7282 16-Core processor, 1.5TB of RAM and 8× NVIDIA A100 GPUs. Python and PyTorch are the main software tools adopted for simulating the federation of clients and their corresponding collaborative training.

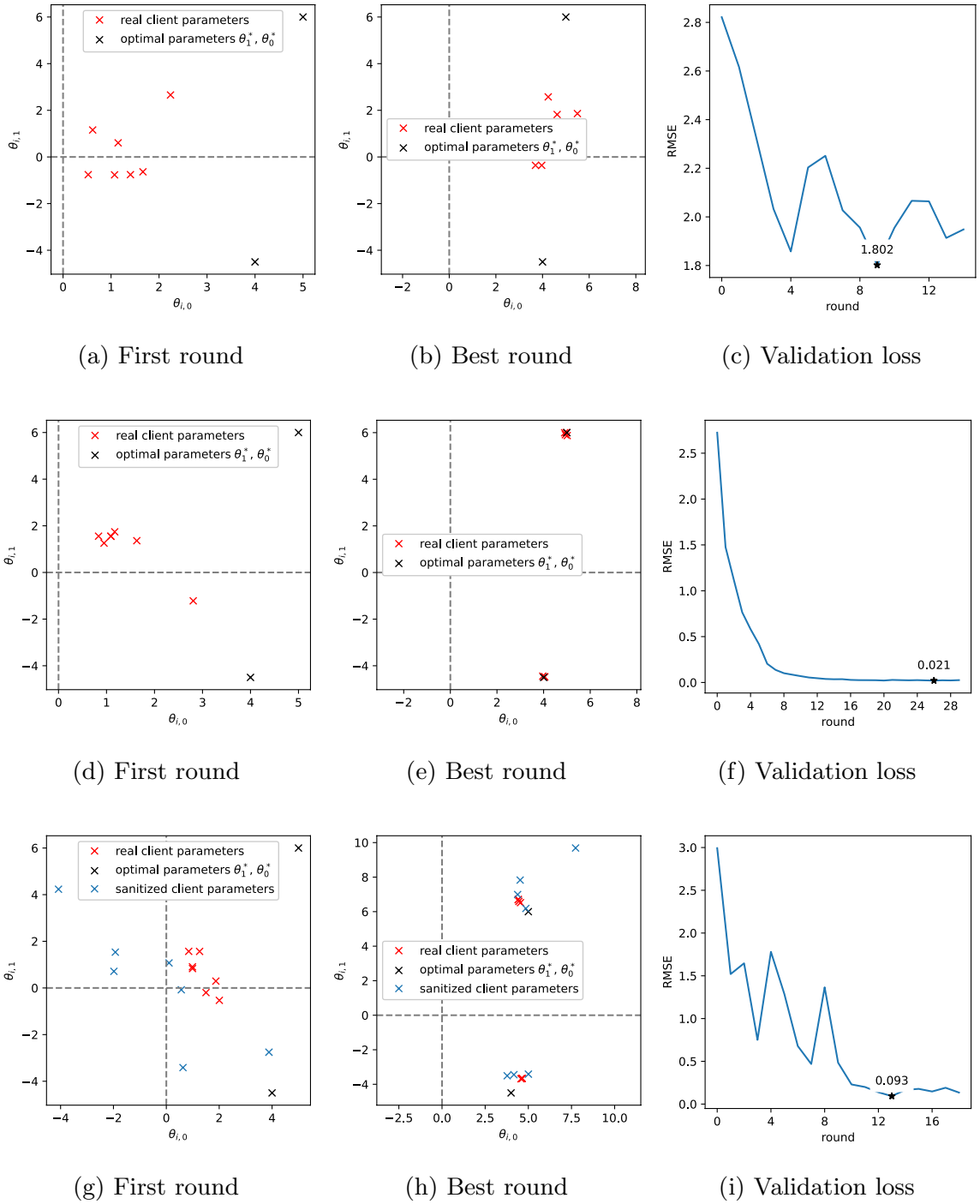
Characterizing Privacy

In this Section, we aim to evaluate and assess the trade-off in training personalized federated learning models under formal local privacy guarantees.

Synthetic Data

Data is generated according to $k = 2$ different distributions: $y = x^T\theta_i^* + u$ and $u \sim \text{Uniform}[0, 1)$, $\forall i \in \{1, 2\}$ and $\theta_1^* = [+5, +6]^T$, $\theta_2^* = [+4, -4.5]^T$. We then assess how training progresses as we move from the Federated Averaging [51] (Fig. 1a–c), to IFCA (Fig. 1d–f), and finally Algorithm 1 (Fig. 1–i). When utilizing Federated Averaging, a noticeable issue arises: relying on a single hypothesis fails to capture the diversity present in the data distributions. As a result, the final parameters tend to settle somewhere between the optimal parameter values (see Fig. 1b). Conversely, employing IFCA demonstrates that having multiple initial hypotheses enhances performance, particularly when clients possess heterogeneous data. This is evident from the nearly overlapping optimized client parameters with the true optimal parameters (see Fig. 1e).

By adopting our algorithm instead, not only do we provide formal guarantees, but we also achieve remarkable outcomes in terms of proximity to the optimal parameters (see Fig. 1h) and reduction of the loss function (see Fig. 1i). To assess privacy infringement, Fig. 2 illustrates the maximum level of privacy leakage incurred by clients per cluster.



(a) First round (b) Best round (c) Validation loss

(d) First round (e) Best round (f) Validation loss

(g) First round (h) Best round (i) Validation loss

Fig. 1 (From [20]) Learning federated linear models with: (**a**–**c**) one initial hypothesis and non-sanitized communication, (**d**–**f**) two initial hypotheses and non-sanitized communication, (**g**–**i**) two initial hypotheses and sanitized communication. The first two figures of each row show the parameter vectors released by the clients to the server

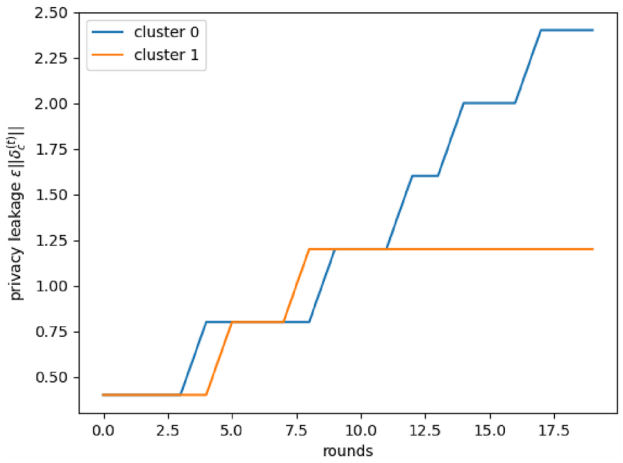


Fig. 2 (From [20]) Synthetic data: max privacy leakage among clients. Privacy leakage is constant when clients with the largest privacy leakage are not sampled (by chance) to participate in those rounds

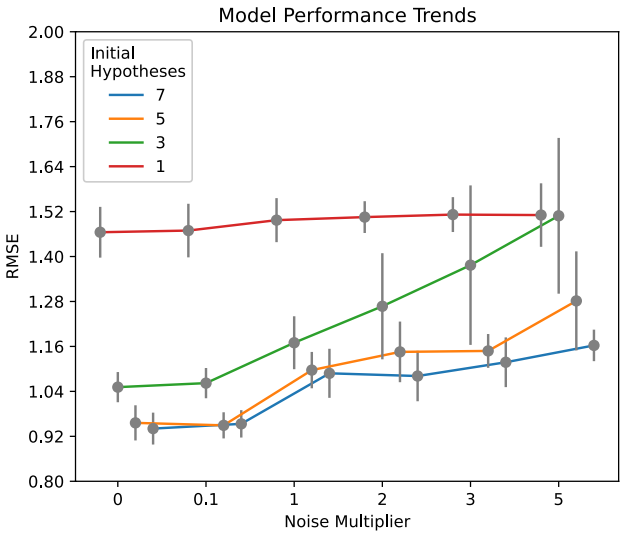


Fig. 3 (From [20]) RMSE for models trained with Algorithm 1 on the Hospital Charge Dataset. Error bars show $\pm\sigma$, with σ the empirical standard deviation. Lower RMSE values are better for accuracy

Hospital Charge Data

This experiment utilizes the Hospital Charge Dataset obtained from the Centers for Medicare and Medicaid Services of the US Government [52]. Here, the healthcare providers are regarded as the clients who participate in training a machine learning model through federated learning. The objective is to predict the cost of a medical service based on its location in the country and the specific procedure involved.

To evaluate the trade-off between privacy, personalization, and accuracy, we explore various numbers of initial

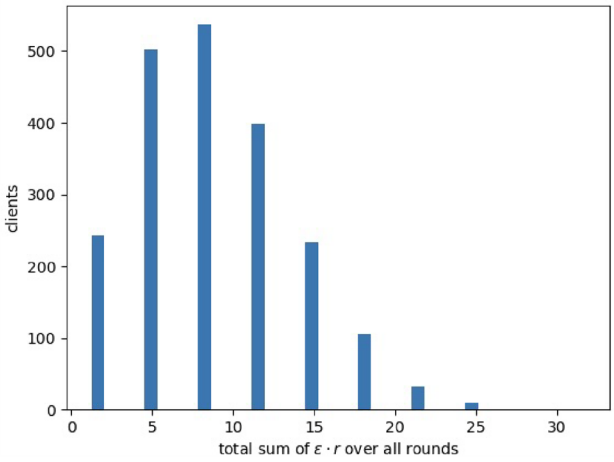


Fig. 4 (From [20]) Hospital charge data: the empirical distribution of the privacy budget over the clients for $\nu = 3$, 5 initial hypotheses, seed = 3, r is the radius of the neighbourhood, the total number of clients is 2062

Table 2 (From [20]) Hospital charge data: median and maximum local privacy budgets over the whole set of clients, averaged over 10 runs with different seeds

	Hypotheses			
ν	7	5	3	1
0	–, –	–, –	–, –	–, –
0.1	517.0, 1551.0	418.0, 1342.0	473.0, 1386.0	528.0, 1540.0
1	36.3, 126.5	40.7, 127.6	44.0, 138.6	49.5, 147.4
2	15.4, 57.8	14.3, 54.5	22.0, 69.3	21.5, 66.6
3	7.7, 32.3	8.4, 36.7	12.5, 40.0	12.1, 40.0
5	5.7, 21.3	5.9, 22.0	5.5, 21.6	5.3, 20.9

$\nu = 0$ means no privacy guarantee

hypotheses since the number of underlying data distributions is unknown a priori. Accuracy is assessed at different levels of the noise multiplier ν. Notably, using Algorithm 1 with only one hypothesis yields the Federated Averaging algorithm. As depicted in Fig. 3, employing multiple hypotheses significantly reduces the RMSE loss function, particularly when transitioning from one to three hypotheses. Furthermore, we emphasize that increasing the number of hypotheses also helps mitigate the impact of the noise multiplier, even at high levels (as shown on the right-hand side of the figure). This highlights the importance of adopting formal privacy guarantees when a slight increase in the cost function is acceptable. The empirical distribution of privacy leakage among clients involved in a specific training configuration is illustrated in Fig. 4. Table 2 presents privacy leakage statistics across multiple rounds and configurations.

Table 3 (From [20]) Effects of increasing the noise multiplier on the validation accuracy and standard deviation

	Cross Entropy loss		RMSE loss	
ν	Average Accuracy	Standard Deviation	Average Accuracy	Standard Deviation
0	0.832	± 0.012	0.801	± 0.001
0.001	0.843	± 0.006	0.813	± 0.014
0.01	0.832	± 0.017	0.805	± 0.008
0.1	0.834	± 0.026	0.808	± 0.019
1	0.834	± 0.014	0.814	± 0.012
3	0.835	± 0.017	0.825	± 0.010
5	0.812	± 0.016	0.787	± 0.003
10	0.692	± 0.002	0.687	± 0.014
15	0.561	± 0.005	0.622	± 0.003

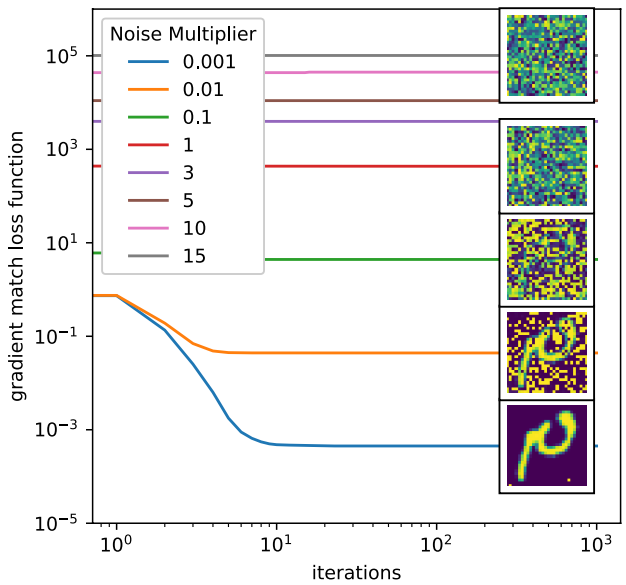


Fig. 5 (From [20]) Effects of the Laplace mechanism in Proposition 4.1 with different noise multipliers as a defence strategy against the DLG attack

FEMNIST Image Classification

This task involves character recognition from images using the FEMNIST dataset [53]. When selecting the range of noise multipliers ν, the resulting privacy leakage $\varepsilon\|\delta_c^{(t)}\|_2 = n/\nu$ would be exceptionally large, given the CNN's $n = 206590$ parameters. Consequently, this renders the mechanism incapable of providing meaningful theoretical privacy guarantees. This issue is commonly encountered with local privacy mechanisms [54], as the expected value of the noise vector's norm, $\mathbb{E}\left[\gamma_{\varepsilon,n}(r)\right]$, exhibits a linear dependence on n: n/ε.

However, it is still possible to evaluate, in practice, whether this specific generalization of the Laplace mechanism can effectively defend against a particular attack known as DLG [9]. The outcomes of varying noise multiplier values are presented in Fig. 5, and Table 3 provides additional details. Notably, when $\nu = 10^{-3}$, the ground truth image can be fully reconstructed. Partial reconstruction remains possible up to $\nu = 10^{-1}$. However, for $\nu \geq 1$, experimental results demonstrate the failure of the DLG attack to reconstruct input samples when the communication between the client and server is protected by the mechanism outlined in Proposition 4.1.

Fairness Analysis

In this section, we analyze how group fairness improves with the personalization of the trained models under d-privacy guarantees when there are two groups with different data distributions. Experiments were performed on synthetic data and the FEMNIST image classification dataset that was used in Section "Characterizing Privacy". To ensure a thorough evaluation, we considered a variety of group fairness metrics in the experiments. In particular, we measured the fairness with respect to equal opportunity [28], equalized odds [28], and demographic parity [27] as explained in Section "Fairness".

In particular, in Figs. 7 and 8, the X-axis denotes the noise multiplier ν representing the amount of d-private noise added to the local updates as explained in Sect. 4.2 and the Y-axis denotes the absolute value of the difference in fairness between the privileged and unprivileged groups with respect to the different metrics of group fairness that we considered.

Synthetic Data

Synthetic data was generated in a method similar to that in Section "Synthetic Data" with the following modifications to enable us to investigate the aspect of group fairness fostered by our method: i) Total number of users is 1000 and each user holds 10 samples. 800 users have data that is generated according to distributions $y = x^T\theta_1 + u$ and $u \sim$ Uniform$[0, 1)$, $\forall i \in \{1, 2\}$, and set as a privileged majority group g_1. The remaining 200 users have data that is generated according to distribution $y = x^T\theta_2 + 15 + u$ and $u \sim$ Uniform$[0, 1)$, $\forall i \in \{1, 2\}$, and set as an unprivileged minority group g_2. In this case, the sensitive attribute considered to evaluate fairness is the group id G where $G \in \{g_1, g_2\}$. ii) For binary classification, we set labels by using the $z = \text{Sigmoid}(Y)$, $\forall y, \hat{y} \in Y$. In the case of g_1, we assign the label 1 if the value of z is greater than or equal to 0.5 and assign the label 0 otherwise. On the other hand, in the case of g_2, the label 1 is assigned when the $z = \text{Sigmoid}(Y - 15)$, $\forall y, \hat{y} \in Y$ is less than or equal to 0.5,

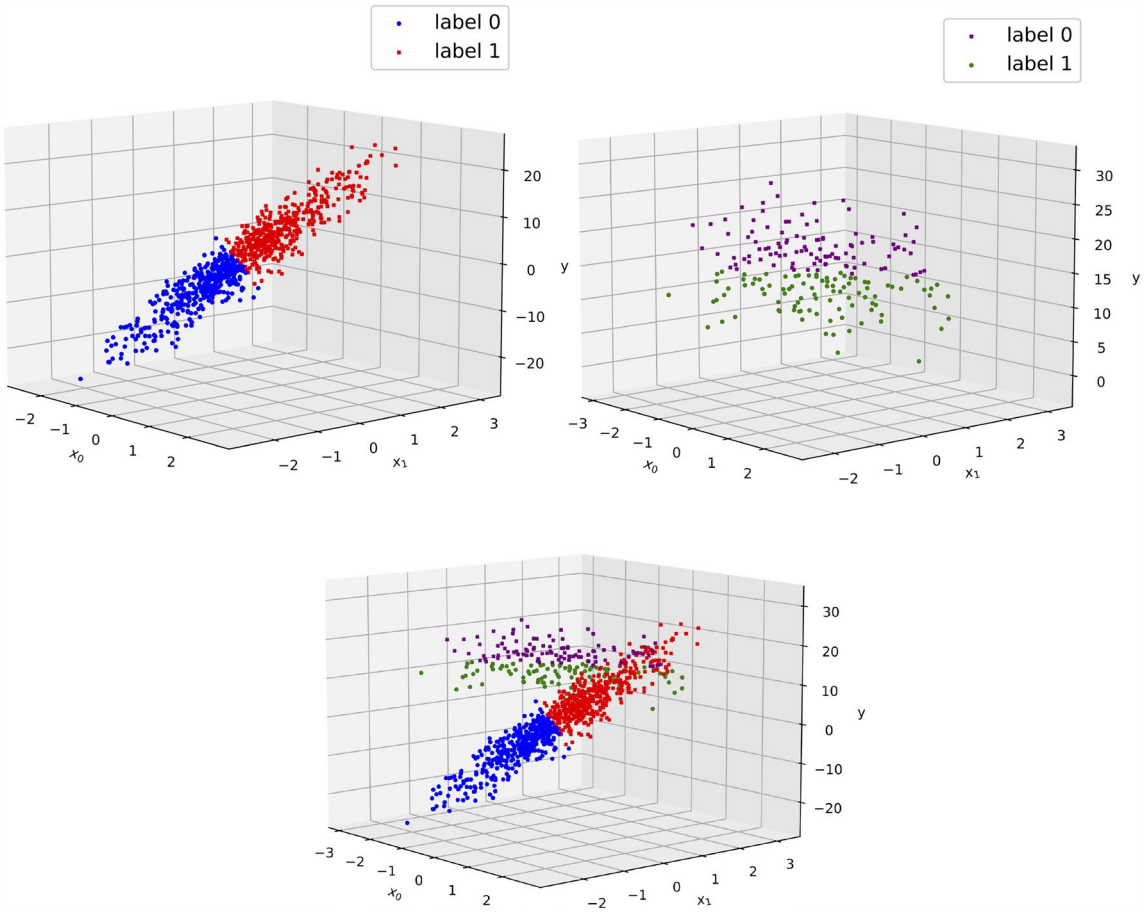
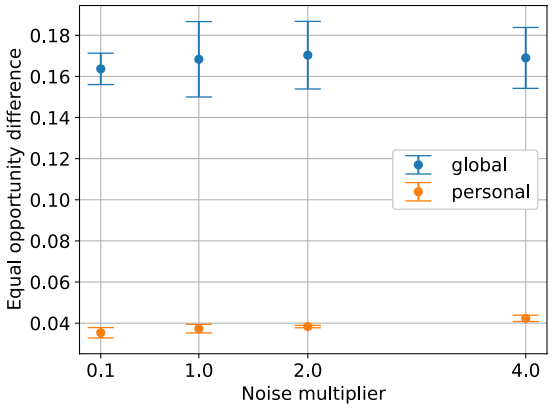


Fig. 6 The first two plots from the left illustrate the spatial distribution of the samples in g_1 and g_2, respectively, and the third plot shows g_1 and g_2 superimposed together in the same space

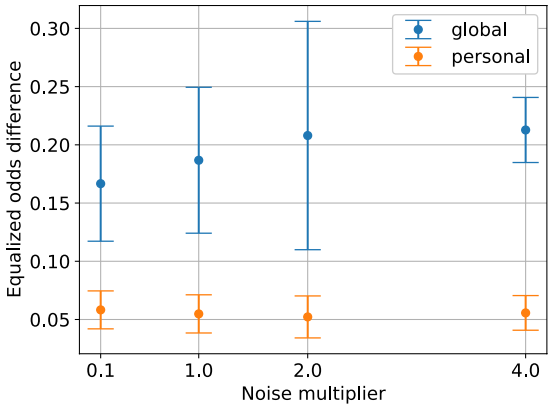
and the label 0 is assigned otherwise. This setting is to simulate a situation in which discrimination occurs depending on sensitive attributes in the real world such as minorities would have experienced a higher loan rejection rate than white applicants with the same property [55]. Thus, in our experiment, label 1 could be interpreted as "loan approved" and label 0 as "loan denied". The data generated in this way are shown in Fig. 6.

We compared the fairness for two cases: one with a single hypothesis (no personalization) and the other with the number of hypotheses as 2 (with personalization) in the framework of Algorithm 1. The experimental results are demonstrated in Fig. 7.

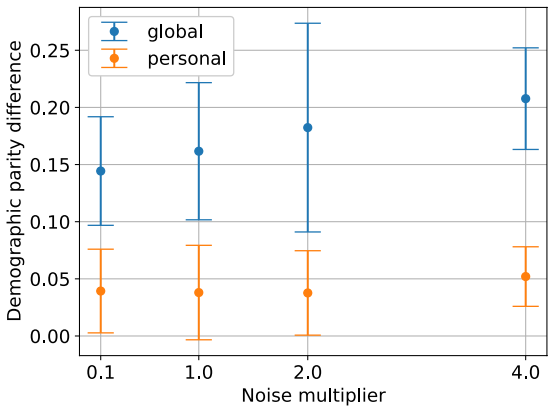
The results illustrated by Fig. 7 assert that the personalization of models (i.e., Algorithm 1) enhances the group fairness under all the metrics and the levels of formal privacy guarantees compared to that of the non-personalized model. A major reason behind this significant improvement of fairness by the personalized model is that unlike the non-personalized model, which trains using data from both groups that are biased towards the majority group g_1, the personalized model training optimizes for each group's data distribution without disregarding the effect of the minority group g_2. We also observe that fairness deteriorates as the value of the noise multiplier increases, as we would expect. This is presumably due to the decreasing influence of the minority group g_2 as the amount of noise insertion increases. This is consistent with the philosophy behind and the definition of DP and its variants. Furthermore, interestingly we observe that the personalized model ensures better fairness than the non-personalized model even with the highest level of privacy protection. This shows that personalization in FL under d-privacy can be a comprehensive solution towards privacy-preserving and ethical machine learning as it provides both privacy guarantees and enhanced fairness.



(a) Equal opportunity difference



(b) Equalized odds difference



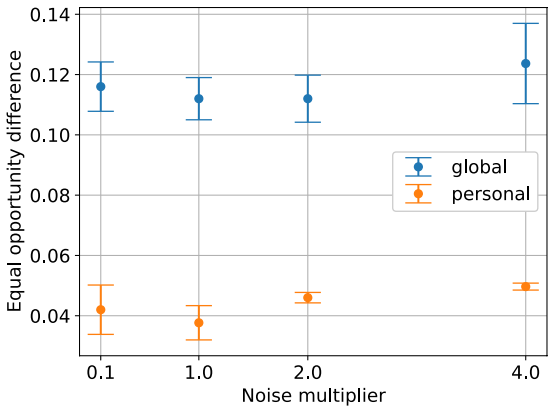
(c) Demographic parity difference

Fig. 7 The figure shows the comparison between the personalized and non-personalized models for (from left) equal opportunity, equalized odds, and demographic parity, respectively. Experiments were performed for noise multipliers ν of 0.1, 1, 2, and 4. For all the metrics of fairness and the values of the noise multiplier, the personalized model is seen to possess improved fairness over the non-personalized model

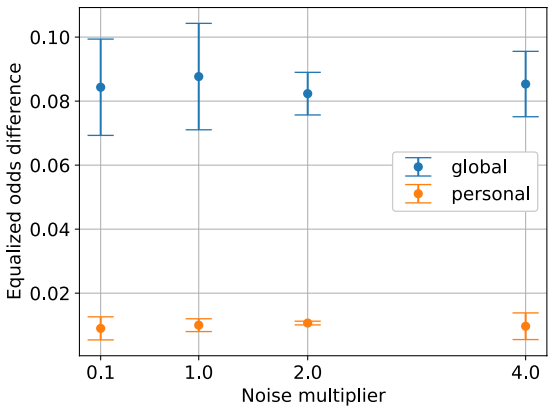
FEMNIST Image Classification

To evaluate the fairness of our method on real datasets, we considered FEMNIST image classification dataset in the same form as in Section "FEMNIST Image Classification". As in experiments performed with the synthetic data in Section "Synthetic data", the size of the groups considered privileged and unprivileged were different denoting the existence of a majority and a minority in the population. In this part, the rotated images are set as the unprivileged group g_2 with a total number of sampled users of 382 forming only 20% of all users. and the un-rotated images are used to represent the privileged group g_1 with a total number of users of 1736. Like in the case of synthetic data considered before, the group membership was used to denote the sensitive attribute. In the case of g_1, we assign label 1 if the FEMNIST image label is even and 0 if it is odd. And for the g_2, we assign label 0 if the FEMNIST image label is even and assign 1 if it is odd. The experimental results are given by Fig. 8.

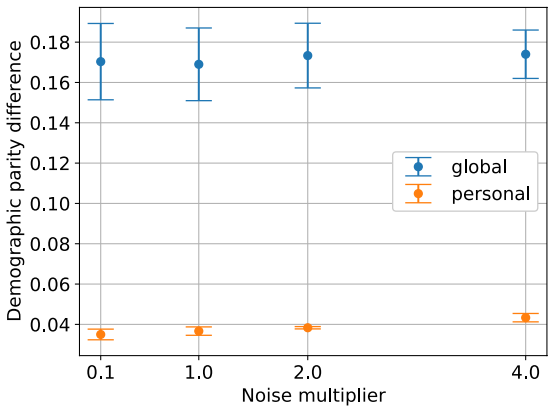
We observe that the personalized model training harbours significantly better group fairness across all metrics



(a) Equal opportunity difference



(b) Equalized odds difference



(c) Demographic parity difference

Fig. 8 The figure shows the comparison between the personalized and non-personalized models for equal opportunity equalized odds, and demographic parity. Experiments were performed for noise multipliers ν of 0.1, 1, 2, and 4. For all metrics of fairness and values of the noise multiplier, the personalized model improved fairness over the non-personalized model

compared to its non-personalized counterpart. The change in fairness due to the amount of noise added was not as notable as in the case of the synthetic dataset but it was still observed to deteriorate with an increase in the value of the noise multiplier. Personalized model training in FL under the highest level of privacy is still observed to have better fairness across all the metrics than (non-personalized) models trained in a classical FL framework even with no privacy, similar to what we observed in the experiments with the synthetic data.

Conclusion

This work builds upon our previous research on personalized federated learning with metric privacy guarantees. To ensure the privacy of ML model parameters during transmission, we employ d-privacy techniques for sanitization. The objective of this process is to generate personalized models that converge to optimal parameters, catering to the diverse datasets present in the federated learning setting. Given the presence of multiple, unknown

data distributions among the individuals participating in the federated learning process, we make a reasonable assumption of a mixture of these distributions. To effectively aggregate clients with similar data distributions, we employ a clustering approach using k-means on the sanitized parameter vectors. This method proves suitable because d-private mechanisms preserve the underlying topology of the true value domain. Notably, our mechanism shows particular promise for machine learning models with a relatively small number of parameters. Although the formal privacy guarantees diminish with larger models, experimental results demonstrate the effectiveness of the Laplace mechanism against the DLG attack.

In addition to metric privacy guarantees, we also evaluate the fairness of machine learning models trained using personalized federated learning and d-privacy. Our study assesses various group fairness metrics, including equal opportunity, equalized odds, and demographic parity. The consistent findings demonstrate that personalized models significantly improve group fairness across all evaluated metrics and privacy levels. Moreover, they, unlike non-personalized models, optimize for each group's specific data distribution, effectively mitigating biases towards the majority group. Consequently, significant advancements in fairness are achieved through this approach.

The level of fairness is influenced by the incorporation of d-private noise in the local updates. As the noise increases, the influence of the minority group decreases, resulting in a deterioration of fairness. This behaviour aligns with the principles of differential privacy and the expected impact of noise addition on group fairness. Remarkably, even with the highest level of privacy protection, personalized models consistently maintain superior fairness compared to non-personalized models. This observation highlights the potential of personalized model training in federated learning under d-privacy as a comprehensive solution for privacy-preserving and ethical machine learning. By offering privacy guarantees alongside enhanced fairness, personalized models demonstrate their effectiveness in balancing these critical aspects.

Funding Open access funding provided by Scuola Normale Superiore within the CRUI-CARE Agreement.

Data availability The FEMNIST dataset that was used to evaluate the effectiveness of our method against DLG attacks and in fairness analysis is available at https://leaf.cmu.edu/. The hospital charge dataset that was used to illustrate the privacy guarantees of our method is available at https://www.cms.gov/mmrr/News/mmrr-news-2013-03-hosp-chg-data.html. The corresponding sections in the paper cite these datasets (item 53 for the FEMNIST dataset and item 52 for the Hospital Charge dataset in the References). The distributions and the heuristics used to generate the synthetic data used to measure the performance of our method for privacy and fairness have been discussed in detail in the corresponding sections.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Le Métayer D, De S.J. PRIAM: a Privacy Risk Analysis Methodology. In: Livraga, G., Torra, V., Aldini, A., Martinelli, F., Suri, N. (eds.) Data Privacy Management and Security Assurance. Springer, Heraklion, Greece 2016. https://hal.inria.fr/hal-01420983
2. NIST: NIST Privacy Framework Core 2021. https://www.nist.gov/system/files/documents/2021/05/05/NIST-Privacy-Framework-V1.0-Core-PDF.pdf
3. McMahan B, Moore E, Ramage D, Hampson S, Arcas B.A. Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics, 2017;pp. 1273–1282. PMLR
4. Ghosh A, Chung J, Yin D, Ramchandran K. An efficient framework for clustered federated learning. Adv Neural Inf Process Syst. 2020;33:19586–97.
5. Mansour Y, Mohri M, Ro J, Suresh A.T. Three approaches for personalization with applications to federated learning. arXiv preprint arXiv:2002.10619 2020.
6. Sattler F, Müller K-R, Samek W. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. IEEE transactions on neural networks and learning systems. 2020;32(8):3710–22.
7. Hitaj B, Ateniese G, Perez-Cruz F. Deep models under the gan: information leakage from collaborative deep learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017;pp. 603–618.
8. Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In: 2019 IEEE Symposium on Security and Privacy (SP), 2019;pp. 739–753. IEEE
9. Zhu L, Liu Z, Han S. Deep leakage from gradients. Advances in Neural Information Processing Systems 2019;**32**.
10. Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: Halevi S, Rabin T, editors.

Theory of Cryptography. Berlin, Heidelberg: Springer; 2006. p. 265–84.
11. Dwork C, Kenthapadi K, McSherry F, Mironov I, Naor M. Our data, ourselves: Privacy via distributed noise generation. In: Vaudenay S, editor. Advances in Cryptology - EUROCRYPT 2006. Berlin, Heidelberg: Springer; 2006. p. 486–503.
12. Andrew G, Thakkar O, McMahan B, Ramaswamy S. Differentially private learning with adaptive clipping. Advances in Neural Information Processing Systems 2021; **34**.
13. McMahan H.B, Ramage D, Talwar K, Zhang L. Learning differentially private recurrent language models. In: International Conference on Learning Representations 2018. https://openreview.net/forum?id=BJ0hF1Z0b
14. Truex S, Liu L, Chow K.-H, Gursoy M.E, Wei W. Ldp-fed: Federated learning with local differential privacy. In: Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking, 2020; pp. 61–66.
15. Zhao Y, Zhao J, Yang M, Wang T, Wang N, Lyu L, Niyato D, Lam K-Y. Local differential privacy-based federated learning for internet of things. IEEE Internet Things J. 2020;8(11):8836–53.
16. Chatzikokolakis K, Andrés M.E, Bordenabe N.E, Palamidessi C. Broadening the scope of differential privacy using metrics. In: International Symposium on Privacy Enhancing Technologies Symposium, 2013; pp. 82–102. Springer
17. Biswas S, Palamidessi C. PRIVIC: A privacy-preserving method for incremental collection of location data 2023.
18. Fernandes N, McIver A, Palamidessi C, Ding M. Universal optimality and robust utility bounds for metric differential privacy. In: 2022 IEEE 35th Computer Security Foundations Symposium (CSF), 2022; pp. 348–363 . https://doi.org/10.1109/CSF54842.2022.9919647
19. Atmaca U.I, Biswas S, Maple C, Palamidessi C. A privacy preserving querying mechanism with high utility for electric vehicles 2022.
20. Galli F, Biswas S, Jung K, Cucinotta T, Palamidessi C. Group Privacy for Personalized Federated Learning. In: Proceedings of the 9th International Conference on Information Systems Security and Privacy - ICISSP, 2023; pp. 252–263 . https://doi.org/10.5220/0011885000003405 . SciTePress - INSTICC
21. Berk R, Heidari H, Jabbari S, Kearns M, Roth A. Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research. 2021;50(1):3–44.
22. Chouldechova A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data. 2017;5(2):153–63.
23. Agarwal S. Trade-Offs between fairness and privacy in machine learning. IJCAI 2021 Workshop on AI for Social Good. 2022; 2021.
24. Verma S, Rubin J. Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness, 2018; pp. 1–7.
25. Hanna R, Linden L. Measuring discrimination in education. National Bureau of Economic Research: Technical report; 2009.
26. Makhlouf K, Zhioua S, Palamidessi C. On the applicability of machine learning fairness notions. ACM SIGKDD Explorations Newsl. 2021;23(1):14–23.
27. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 2012; pp. 214–226.
28. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. Advances in neural information processing systems 2016; **29**.
29. Abadi M, Chu A, Goodfellow I, McMahan H.B, Mironov I, Talwar K, Zhang L. Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016; pp. 308–318.
30. Geyer R.C, Klein T, Nabi M. Differentially private federated learning: A client level perspective. arXiv preprint arXiv:1712.07557 2017.
31. Bonawitz K.A, Ivanov V, Kreuter B, Marcedone A, McMahan H.B, Patel S, Ramage D, Segal A, Seth K. Practical secure aggregation for federated learning on user-held data. In: NIPS Workshop on Private Multi-Party Machine Learning 2016 . https://arxiv.org/abs/1611.04482
32. Agarwal N, Suresh A.T, Yu F.X.X, Kumar S, McMahan B. cpsgd: Communication-efficient and differentially-private distributed sgd. Advances in Neural Information Processing Systems 2018; **31**.
33. Hu R, Guo Y, Li H, Pei Q, Gong Y. Personalized federated learning with differential privacy. IEEE Internet Things J. 2020;7(10):9530–9.
34. Bonawitz K.A, Ivanov V, Kreuter B, Marcedone A, McMahan H.B, Patel S, Ramage D, Segal A, Seth K. Practical secure aggregation for federated learning on user-held data. In: NIPS Workshop on Private Multi-Party Machine Learning 2016. https://arxiv.org/abs/1611.04482
35. Chhabra S, Solihin Y, Lal R, Hoekstra M. An analysis of secure processor architectures. Transactions on computational science VII, 2010; 101–121.
36. Cucinotta T, Cherubini D, Jul E. Confidential execution of cloud services. In: CLOSER, 2014; pp. 616–621.
37. Chhabra A, Masalkovaitė K, Mohapatra P. An overview of fairness in clustering. IEEE Access. 2021;9:130698–720.
38. Ezzeldin Y.H, Yan S, He C, Ferrara E, Avestimehr S. Fairfed: Enabling group fairness in federated learning. In: 1st NeurIPS Workshop on New Frontiers in Federated Learning 2021. https://arxiv.org/abs/1611.04482
39. Chu L, Wang L, Dong Y, Pei J, Zhou Z, Zhang Y. Fedfair: Training fair models in cross-silo federated learning. arXiv preprint arXiv:2109.05662 2021.
40. Menon A.K, Williamson R.C. The cost of fairness in binary classification. In: Conference on Fairness, Accountability and Transparency, 2018; pp. 107–118 . PMLR
41. Wick M, Tristan J.-B, et al. Unlocking fairness: a trade-off revisited. Advances in neural information processing systems 2019; **32**.
42. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR). 2021;54(6):1–35.
43. Biswas S, Rajan H. Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline. In: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2021; pp. 981–993.
44. Kamiran F, Calders T. Data preprocessing techniques for classification without discrimination. Knowl Inf Syst. 2012;33(1):1–33.
45. Wan M, Zha D, Liu N, Zou N. In-processing modeling techniques for machine learning fairness: A survey. ACM Trans Knowl Discov Data. 2023;17(3):1–27.
46. Hashimoto T, Srivastava M, Namkoong H, Liang P. Fairness without demographics in repeated loss minimization. In: International Conference on Machine Learning, 2018; pp. 1929–1938 . PMLR

47. Petersen F, Mukherjee D, Sun Y, Yurochkin M. Post-processing for individual fairness. Adv Neural Inf Process Syst. 2021;34:25944–55.
48. Noriega-Campero A, Bakker M.A, Garcia-Bulle B, Pentland A. Active fairness in algorithmic decision making. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019; pp. 77–83.
49. Cummings R, Gupta V, Kimpara D, Morgenstern J. On the compatibility of privacy and fairness. In: Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, 2019; pp. 309–315.
50. Andrés M.E, Bordenabe N.E, Chatzikokolakis K, Palamidessi C. Geo-indistinguishability: Differential privacy for location-based systems. In: Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security, 2013; pp. 901–914.
51. Konečný J, McMahan H.B, Yu F.X, Richtarik P, Suresh A.T, Bacon D. Federated learning: Strategies for improving communication efficiency. In: NIPS Workshop on Private Multi-Party Machine Learning 2016. https://arxiv.org/abs/1610.05492
52. CMMS: Centers for Medicare and Medicaid Services. Accessed: 2021; 2022-09-21 . https://www.cms.gov/mmrr/News/mmrr-news-2013-03-hosp-chg-data.html
53. Caldas S, Duddu S.M.K, Wu P, Li T, Konečnỳ J, McMahan H.B, Smith V, Talwalkar A. Leaf: A benchmark for federated settings. Workshop on Federated Learning for Data Privacy and Confidentiality 2019.
54. Bassily R, Nissim K, Stemmer U, Guha Thakurta A. Practical locally private heavy hitters. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc., Red Hook, NY, USA 2017. https://proceedings.neurips.cc/paper/2017/file/3d779cae2d46cf6a8a99a35ba4167977-Paper.pdf
55. Bartlett R, Morse A, Stanton R, Wallace N. Consumer-lending discrimination in the fintech era. J Financ Econ. 2022;143(1):30–56.