SHARE YOUR REPRESENTATION ONLY: GUARANTEED IMPROVEMENT OF THE PRIVACY-UTILITY TRADEOFF IN FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Repeated parameter sharing in federated learning causes significant information leakage about private data, thus defeating its main purpose: data privacy. Mitigating the risk of this information leakage, using state of the art differentially private algorithms, also does not come for free. Randomized mechanisms can prevent convergence of models on learning even the useful representation functions, especially if there is more disagreement between local models on the classification functions (due to data heterogeneity). In this paper, we consider a representation federated learning objective that encourages various parties to collaboratively refine the consensus part of the model, with differential privacy guarantees, while separately allowing sufficient freedom for local personalization (without releasing it). We prove that in the linear representation setting, while the objective is non-convex, `CENTAUR` converges to a ball centered around the *global optimal* solution at a linear rate, and the radius of the ball is proportional to the reciprocal of the privacy budget. With this novel utility analysis, we improve the SOTA utility-privacy trade-off for this problem by a factor of $\sqrt{d}$, where d is the input dimension. We empirically evaluate our method with the image classification task on CIFAR10, CIFAR100, and EMNIST, and observe a significant performance improvement over the prior work under the same small privacy budget.

1 INTRODUCTION

In federated learning (FL), multiple parties cooperate to learn a model under the orchestration of a central server while keeping the data local. However, this paradigm alone is insufficient to provide rigorous privacy guarantees, even when local parties only share partial information (e.g. gradients) about their data. An adversary (e.g. one of the parties) can infer whether a particular record is in the training data set of other parties (Nasr et al., 2019), or even precisely reconstruct their training data (Zhu et al., 2019). To formally mitigate these privacy risks, we need to guarantee that any information *shared between the parties* during the training phase has *bounded information leakage* about the local data. This can be achieved using FL under differential privacy (DP) guarantees.

FL and DP are relatively well-studied separately. However, their challenges multiply when conducting FL under a DP constraint, in real-world settings where the data distributions can vary substantially across the clients (Li et al., 2020b; Acar et al., 2021; Shen et al., 2022). A direct consequence of such data heterogeneity is that the optimal local models might vary significantly across clients and differ drastically from the global solution. This results in large local gradients (Jiang et al., 2019). However, these large signals leak information about the local training data, and cannot be communicated as such when we need to guarantee DP. We require clipping gradient values (usually by a small threshold (De et al., 2022)) before sending them to the server, to bound the sensitivity of the gradient function with respect to changes in training data Abadi et al. (2016). As the local per-sample gradients (due to data heterogeneity) tend to be large even at the global optimum, clipping per-example gradient by a small threshold and then randomizing it, will result in a high error in the overall gradient computation, and thus degrading the accuracy of the model learned via FL.

Contributions. In order to achieve high-accuracy FL under a tight privacy budget, we propose a novel framework that is based on learning information that is agreed on among most par-

ties in a differentially private manner, while also allowing each client to personalize its learning on information that others disagree with. It is shown that in many FL classification scenarios, participants have minimal disagreement on data representations (Bengio et al., 2013; Chen et al., 2020; Collins et al., 2021), but possibly have very different classifier heads (e.g., the last layer of the neural network). Therefore, instead of solving the standard classification problem, we view the neural network model as a composition of a representation extractor and a small classifier head, and optimize these two components in different manners. In our scheme, we train a single differentially private global representation extractor while allowing each participant to have a different *personalized* classifier head. Due to low heterogeneity in data representation (compared to the whole model), the DP learned representation in our new scheme outperforms prior schemes that perform DP optimization over the entire model. In the setting where both the representation function and the classifier heads are linear w.r.t. their parameters, we prove a novel utility-privacy trade-off for an instance of CENTAUR, yielding a significant $O(\sqrt{d})$ improvement over previous art, where d is the input dimension (Corollary 5.1). A major algorithmic novelty of our proposed approach is a cross-validation scheme for boosting the success probability of the classic noisy power method for privacy-preserving spectral analysis.

We present strong empirical evidence for the superior performance of CENTAUR over the prior work, under the small DP budget of $(1, 10^{-5})$ in a variety of data-heterogeneity settings on benchmark datasets CIFAR10, CIFAR100, and EMNIST. Our method outperforms the prior work in all settings. Moreover, we showcase that CENTAUR *uniformly* enjoys a better utility-privacy trade-off over its competitors on the EMNIST dataset across different privacy budget ϵ (Figure 1). Importantly, CENTAUR ($\epsilon > 0.5$) is **the only method** that could **outperform the local stand-alone training baseline**, and therefore justifies the choice of collaborative learning as opposed to stand-alone training.

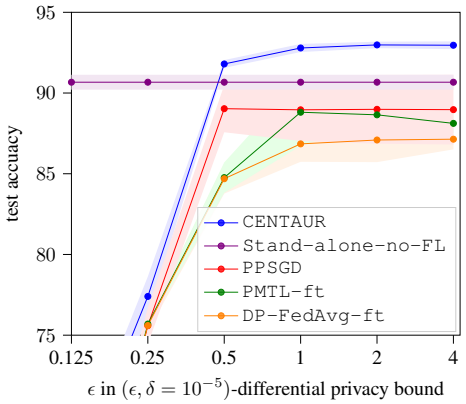

Figure 1: Privacy utility trade-off for models trained under CENTAUR and other algorithms on EMNIST dataset (2000 clients, 5 shards per user). Error bar denotes standard deviation across 3 runs of the algorithms.

1.1 Related Work

Federated learning with differential privacy (DP) guarantee has been extensively studied since its emergence (Shokri & Shmatikov, 2015; McMahan et al., 2017a). Without assuming any trusted central party, the model of local DP requires each client to randomize its messages before sending them to any other (malicious) parties. Consequently, the trade-off between local DP and accuracy is significantly worse than that for centralized setting and requires huge amount of data for learning even simple statistics (Duchi et al., 2014; Erlingsson et al., 2014; Ding et al., 2017). By using secure aggregation protocol as an equivalent for trusted central party, recent works (McMahan et al., 2017b; Agarwal et al., 2018; Levy et al., 2021; Kairouz et al., 2021) study user-level DP under Billboard model to enable utility. We also focus on such user-level DP setting.

Model personalization approaches (Smith et al., 2017; Fallah et al., 2020; Li et al., 2020b; Arivazhagan et al., 2019; Collins et al., 2021; Pillutla et al., 2022) enable each client to learn a different (while related) model, thus alleviating the model drifting issue due to data heterogeneity. Recent works further investigate whether model personalization approaches enable improved privacy accuracy trade-off for federated learning. Hu et al. (2021) propose a private federated multi-task learning algorithm by adding task-specific regularization to each client's optimization objective. Hence, the algorithm is similar to the regularized FedAvg-ft algorithm, and has limited ability to deal with data heterogeneity. Bietti et al. (2022) propose the PPSGD algorithm that enables training additive personalized models with user-level DP. However, their generalization guarantees crucially rely on the convexity of loss functions. On the contrary, we study the convergence of CENTAUR algorithm under more general non-convex objectives.

The closest work to our approach in the literature is Jain et al. (2021), who also propose differentially privately learning shared low-dimensional linear representation with individualized classification head. However, their algorithm relies on solving the least square problem exactly at server side,

by performing SVD on perturbed noisy representation matrix. Hence, the generalization guarantee of their algorithm intrinsically has an expensive $d^{1.5}$ dependency on the data dimension d. By contrast, we perform noisy gradient descent at server side and improve upon this error dependency by a factor of $\sqrt{d}$. Their algorithm is also limited to the linear representation learning problem, unlike our `CENTAUR` algorithm which enables training multiple layers of shared representations.

Our work builds on the `FedRep` algorithm (Collins et al., 2021), which also relies on learning shared representations between clients but does not consider privacy. In contrast, our work provides a novel private and federated representation learning framework. Moreover, a major different ingredient of our algorithm is the initialization procedure, which requires performing *differentially private* SVD to the data matrix. We use the noisy power method (Hardt & Price, 2014) as a crucial tool to enable a constant probability for utility guarantee. We then perform cross-validation to further boost success probability to arbitrarily large (inspired by Liang et al. (2014)).

2 Notations and Background on Privacy

Notations. We adhere to the notations suggested by the ICLR template, with the following additional ones. We denote the clipping operation by $\mathrm{clip}(\boldsymbol{x};\zeta) \doteq \boldsymbol{x} \cdot \min\{1, \zeta/\|\boldsymbol{x}\|\}$. We denote the Gaussian mechanism as $\mathrm{GM}_{\zeta,\sigma}(\{\boldsymbol{x}_i\}_{i=1}^s) \doteq \frac{1}{s}(\sum_{i=1}^s \mathrm{clip}(\boldsymbol{x}_i;\zeta) + \sigma\zeta W)$ where $W \sim \mathcal{N}(0, \mathbb{I})$.

Define Rényi Differential Privacy (RDP) on a dataset space $\mathcal{D}$ equipped with a distance d as follows.

Definition 2.1 (Rényi Differential Privacy (RDP))**.** *For a pair of measures ν and ν' over the same space with $\nu \ll \nu'$, the Rényi divergence $R_\alpha(\nu, \nu')$ is defined as*

$$R_\alpha(\nu,\nu') = \frac{1}{\alpha-1}\log E_\alpha(\nu,\nu'), \text{ where } E_\alpha(\nu,\nu') = \int \left(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\right)^\alpha \mathrm{d}\nu'.$$

Let Θ be an output space. A randomized algorithm $\mathcal{M} : \mathcal{D} \to \Theta$ satisfies (α, ϵ)-RDP, if for all $\mathbb{D}, \mathbb{D}' \in \mathcal{D}$ with $\mathrm{d}(\mathbb{D}, \mathbb{D}') \le 1$, we have $R_\alpha(\mathcal{M}(\mathbb{D}), \mathcal{M}(\mathbb{D}')) \le \epsilon$.

User-level-RDP. Let $\mathcal{D}$ be the space of all of n tuples of local datasets $\{\mathbb{S}_i\}_{i=1}^n$, where each local dataset consists of m data points, i.e. $\mathcal{D} = \{\{\mathbb{S}_i\}_{i=1}^n \mid \mathbb{S}_i = \{\boldsymbol{z}_{ij}\}_{j=1}^m\}$. The distance d is the Hamming distance in the dataset level, i.e. $\mathrm{d}(\{\mathbb{S}_i\}_{i=1}^n, \{\mathbb{S}'_i\}_{i=1}^n) = \sum_{i=1}^n \mathbf{1}_{\mathbb{S}_i \neq \mathbb{S}'_i}$. We refer to the privacy guarantee recovered by this choice of dataset space as user-level-RDP.

Our privacy analysis relies on the Gaussian Mechanism and the Composition of RDP which we review in Appendix A.3. To facilitate the comparison of our result with previous works, we define the standard notion of Differential Privacy (DP) and the billboard model. Note that one can convert RDP to DP via the conversion lemma (see Appendix A.3).

Definition 2.2 (Differential Privacy (DP))**.** *Let Θ be an abstract output space. A randomized algorithm $\mathcal{M} : \mathcal{D} \to \Theta$ is (ϵ, δ)-differential private if for all $\mathbb{D}, \mathbb{D}' \in \mathcal{D}$ with $\mathrm{d}(\mathbb{D}, \mathbb{D}') \le 1$, we have that for all subset of the range, $S \subseteq \Theta$, the algorithm $\mathcal{M}$ satisfies:*

$$\Pr\{\mathcal{M}(\mathbb{D}) \in S\} \le \exp(\epsilon)\Pr\{\mathcal{M}(\mathbb{D}') \in S\} + \delta.$$

The Billboard model is a communication protocol that is particularly compatible with algorithms in the federated setting and has been adopted in many previous works (Jain et al., 2021; Hu et al., 2021; Bietti et al., 2022). In this model, a trusted server (either a trusted party or one that uses cryptographic techniques like multiparty computation) aggregates information subject to a DP constraint, which is then shared as public messages with all the n users. Then, each user computes its own personalized model based on the public messages and its own private data (Hsu et al., 2014).

Threat Models. In this work, we aim to ensure the privacy protection against potential adversarial clients so that any eavesdropper will not be able to tell whether one of the users has participated in the collaborative learning procedure, given the information released during the training phase. Consequently, we will establish user-level RDP guarantees under the billboard model, which is a setting commonly considered in the literature (Jain et al., 2021; Bietti et al., 2022; Hu et al., 2021).

3 Problem Formulation

Consider a Federated Learning (FL) setting where there are n agents interacting with a central server and the ith agent holds m_i local data points $\mathbb{S}_i = \{(\boldsymbol{x}_{ij}, y_{ij})\}_{j=1}^{m_i}$. Here $(\boldsymbol{x}_{ij}, y_{ij}) \in \mathbb{R}^d \times \mathbb{R}$ denotes

a pair of an input vector and the corresponding label. Suppose that each local dataset $\mathbb{S}_i$ is sampled from an underlying joint distribution $p_i(\boldsymbol{x}, y)$ of the input and response pair. We consider the *data-heterogenous* setting where potentially $p_i \neq p_j$ for $i \neq j$. We further consider a hypothesis model $f(\cdot; \boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R}$ which maps an input $\boldsymbol{x}$ to the predicted label y. Here $\boldsymbol{\theta} \in \mathbb{R}^q$ is the trainable parameter of the model f. Let $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a loss function that penalizes the error of the prediction $f(\boldsymbol{x}; \boldsymbol{\theta})$ given the true label y. The local objective on client i is defined as

$$l(\boldsymbol{\theta}; \mathbb{S}_i) \doteq \frac{1}{m_i} \sum_{j=1}^{m_i} \ell(f(\boldsymbol{x}_{ij}; \boldsymbol{\theta}), y_{ij}). \tag{1}$$

The standard FL formulation seeks a global consensus solution, whose objective is defined as

$$\arg\min_{\theta} \mathcal{L}(\boldsymbol{\theta}) \doteq \frac{1}{n} \sum_{i=1}^{n} l(\boldsymbol{\theta}; \mathbb{S}_i). \tag{2}$$

While this formulation is reasonable when the local data distributions are similar, the obtained global solution may be far from the local optima $\arg\min l(\boldsymbol{\theta}; \mathbb{S}_i)$ under diverse local data distributions, a phenomenon known as *statistical heterogeneity* in the FL literature (Li et al., 2020a; Wang et al., 2019; Malinovskiy et al., 2020; Mitra et al., 2021; Charles & Konečnỳ, 2021; Acar et al., 2020; Karimireddy et al., 2020). Such a (potentially significant) mismatch between local and global optimal solutions limits the incentive for collaboration, and cause extra difficulties when DP constraints are imposed (Remark 3.1). These considerations motivate us to search for personalized solutions that can be learned in a federated fashion, with less utility loss due to the DP constraint.

Federated representation learning with differential privacy. It is now well-documented that in some common and real-world FL tasks, such as image classification and word prediction, clients have minimal disagreement on data representations (Chen et al., 2020; Collins et al., 2021). Based on this observation, a more reasonable alternative to the FL objective in equation 2 should focus on learning the data representation, which is the information that is agreed on among most parties, while also allowing each client to personalize its learning on information that the other clients disagree on. To formalize this, suppose that the variable $\boldsymbol{\theta} \in \mathbb{R}^q$ can be partitioned into a pair $[\boldsymbol{w}, \boldsymbol{b}] \in \mathbb{R}^{q_1} \times \mathbb{R}^{q_2}$ with $q = q_1 + q_2$ and the parameterized model admits the composition $f = h \circ \phi$ where $\phi : \mathbb{R}^d \to \mathbb{R}^k$ is the *representation extractor* that maps d-dimensional data points to a lower dimensional space of size k and $h : \mathbb{R}^k \to \mathbb{R}$ is a *classifier head* that maps from the lower dimensional subspace to the space of labels. An important example is bottom and top layers of the neural network model. We use $\boldsymbol{w}$ and $\boldsymbol{b}$ to denote the parameters that determine h and ϕ, respectively. With the above notation, we consider the following federated representation learning (FRL) problem

$$\min_{\boldsymbol{b} \in \mathbb{B}} \frac{1}{n} \sum_{i=1}^{n} \min_{\boldsymbol{w}_i} \left\{ l([\boldsymbol{w}_i, \boldsymbol{b}]; \mathbb{S}_i) := \frac{1}{m_i} \sum_{j=1}^{m_i} \ell(h(\phi(\boldsymbol{x}_{ij}; \boldsymbol{b}); \boldsymbol{w}_i), y_{ij}) \right\}, \tag{3}$$

where we maintain a single global representation extraction function $\phi(\cdot; \boldsymbol{b})$ subject to the constraint $\boldsymbol{b} \in \mathbb{B} \subseteq \mathbb{R}^{q_2}$ while allowing each client to use its personalized classification head $h(\cdot; \boldsymbol{w}_i)$ locally. The constraint $\mathbb{B}$ is included so that equation 3 also covers the linear case studied in section 5.

The choice of the FRL formulation in equation 3 entails considerations from both DP and optimization perspectives: From the DP standpoint, the phenomenon of statistical heterogeneity introduces additional difficulties for federated learning under DP constraint (see Remark 3.1 below). If the clients collaborate to train only a shared representation function, then the aforementioned disadvantages can be alleviated; From the optimization standpoint, we typically have $k \ll d$, i.e. the dimension of the extracted features is much smaller than that of the original input. Hence, for a fixed representation function $\phi(\cdot; \boldsymbol{b})$, the client specific heads $h(\cdot; \boldsymbol{w}_i)$ are in general easy to optimize locally as the number of parameters, k, is typically small.

Remark 3.1 (Statistical heterogeneity makes DP guarantee harder to establish.)**.** *To establish DP guarantees for gradient based methods, e.g. DP-SGD, a common choice is the Gaussian mechanism, which is comprised of the gradient clipping step and the noise injection step. It is empirically observed that to achieve a better privacy-utility trade-off, a small clipping threshold is preferred, since it limits the large variance due to the injected noise (De et al., 2022). Moreover, the effect of the bias (due to clipping) subsides as the per-sample gradient norm diminishes during the centralized*

Algorithm 1 SERVER procedure of CENTAUR

procedure SERVER($\boldsymbol{b}^0$, p_g, T_g, η_g, σ_g, ζ_g)
  // $\boldsymbol{b}^0$ is obtained from the INITIALIZATION procedure.
  for $t \leftarrow 0$ to $T_g - 1$ **do**
    Sample set $\mathbb{C}^t$ of active clients using Poisson sampling with parameter p_c.
    Broadcast the current global representation parameter $\boldsymbol{b}^t$ to active clients.
    Receive the local update directions $\{\boldsymbol{g}_i^t\}_{i\in\mathbb{C}^t}$ from the CLIENT procedures.
    Compute the update direction $\boldsymbol{g}^t = \mathrm{GM}_{\zeta_g,\sigma_g}(\{\boldsymbol{g}_i^t\}_{i\in\mathbb{C}^t})$
    Update the global representation function $\boldsymbol{b}^{t+1} :=$ AGGREGATION($\boldsymbol{b}^t, \boldsymbol{g}^t, \eta_g$).
    // The AGGREGATION procedure depends on the feasible set $\mathbb{B}$ in equation 3.

Algorithm 2 CLIENT procedure of CENTAUR in the general case (for client i)

procedure CLIENT($\boldsymbol{b}^t$, $\bar{m}$, T_l, η_l)
  [Phase 1: Local classifier update.] $\boldsymbol{w}_i^{t+1} = \arg\min_{\boldsymbol{w}} l([\boldsymbol{b}^t, \boldsymbol{w}]; \mathbb{S}_i)$.
  [Phase 2: Local representation function update.] Set $\boldsymbol{b}_i^{t,0} = \boldsymbol{b}^t$;
  for $s \leftarrow 0$ to $T_l - 1$ **do**
    Sample a subset $\mathbb{S}_i^s$ of size $\bar{m}$ from the local dataset $\mathbb{S}_i$ without replacement
    Update the local representation function $\boldsymbol{b}_i^{t,s+1} := \boldsymbol{b}_i^{t,s} - \eta_l \cdot \partial_{\boldsymbol{b}} l([\boldsymbol{b}_i^{t,s}, \boldsymbol{w}_i^{t+1}]; \mathbb{S}_i^s)$.
  [Phase 3: Summarize the local update direction.] **return** $\boldsymbol{g}_i^t := \boldsymbol{b}_i^{t,T_l} - \boldsymbol{b}^t$.

training, a phenomenon known as benign overfitting in deep learning (Bartlett et al., 2020; Li et al., 2021; Bartlett et al., 2021). However, due to the phenomenon of distribution shift, the local (per-sample) gradients in the standard FL setting (described in equation 2) remain large even at the global optimal solution, and hence setting a small (per-sample) gradient clipping threshold will result in a large and non-diminishing bias in the overall gradient computation.
In contrast, for tasks where the representation extracting functions are approximately homogeneous, the local and global optimal of the FRL formulation 3 are close and hence the gradients w.r.t. the representation function vanishes at the optimal, which amiable to small clipping threshold.

4 DIFFERENTIAL PRIVATE FEDERATED REPRESENTATION LEARNING

In this section we present the proposed CENTAUR method to solve the FRL problem in equation 3.

SERVER procedure (Algorithm 1) takes the following quantities as inputs: $\boldsymbol{b}^0$ denotes the initializer for the parameter of the global representation function, obtained from a procedure INITIALIZATION; p_g denotes the portion of the clients that will participate in training per global communication round; T_g denotes the total number of global communication rounds; η_g denotes the global update step size; (σ_g, ζ_g) stand for the noise multiplier and the clipping threshold of the Gaussian mechanism (that ensures user-level RDP). Note that in this section, we consider random INITIALIZATION over unconstrained space $\mathbb{B} = \mathbb{R}^{q_2}$, and the procedure AGGREGATION($\boldsymbol{b}^t, \boldsymbol{g}^t, \eta_g$) $= \boldsymbol{b}^t + \eta_g \cdot \boldsymbol{g}^t$. Under these configurations, SERVER follows the standard FL protocol: After broadcasting the current global representation function to the activate clients, it aggregates the information returned from the CLIENT procedure to update the global representation function.

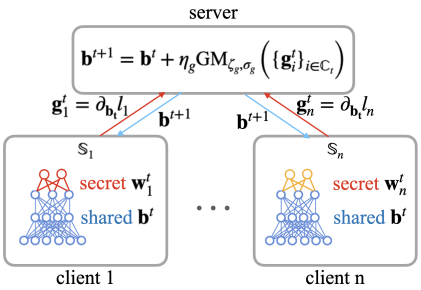

Figure 2: The t-th global round of the CENTAUR algorithm, where clients keep their classification head $\boldsymbol{w}_i^t$ secret while updating shared representation $\boldsymbol{b}^t \to \boldsymbol{b}^{t+1}$ based on perturbed gradients $\boldsymbol{g}_i^t$ from sampled clients $i \in \mathbb{C}_t$.

CLIENT procedure (Algorithm 2) takes the following quantities as inputs: $\boldsymbol{b}^t$ denotes the parameter of the global representation function received from the server; $\bar{m}$ denotes the number of local data points used as mini-batch to update the local representation function; T_l denotes the number of local update iterations; η_l denotes the local update step size. CLIENT can be divided

into three phases: 1. After receiving the current global parameter $\boldsymbol{b}^t$ of the representation function from the server, the client update the local classifier head to the minimizer of the local objective $\boldsymbol{w}_i^{t+1} = \arg\min_{\boldsymbol{w}} l([\boldsymbol{b}^t, \boldsymbol{w}]; \mathbb{S}_i)$. This is possible since the local objective l usually admits very simple structure, e.g. it is convex w.r.t. $\boldsymbol{w}$, once the representation function is fixed. In practice, we would run multiple SGD epochs on $\boldsymbol{w}$ to approximately minimize l with $\boldsymbol{b}^t$ fixed. This is computationally cheap since the dimension of $\boldsymbol{w}$ is much smaller compared to the whole variable $\boldsymbol{\theta} = [\boldsymbol{w}, \boldsymbol{b}]$. 2. Once the local classifier head is updated to $\boldsymbol{w}_i^{t+1}$, the client optimizes its local representation function with multiple SGD steps, starting from the current global consensus $\boldsymbol{b}^t$. 3. Finally, each client calculate its local update direction for representation via the difference between its latest local representation function and the previous global consensus, $\boldsymbol{b}_i^{t,T_l} - \boldsymbol{b}^t$.

Remark 4.1 (Privacy guarantee)**.** *By the composition theorem for subsampled Gaussian Mechanism Mironov et al. (2019), we prove that Algorithm 1 and Algorithm 2 satisfies user-level (α, ϵ)-Rényi differential privacy for $\epsilon = T_g \cdot S_\alpha(p_c, \sigma_g)$, where $S_\alpha(p_c, \sigma_g) = R_\alpha(\mathcal{N}(0, \sigma_g^2) \| \mathcal{N}(0, \sigma_g^2) + p_c \cdot \mathcal{N}(1, \sigma_g^2))$. In the special case of full gradient descent with $p_c = 1$, we have that $\epsilon = \alpha \cdot T_g/(2\sigma_g^2)$.*

5 Guaranteed Improvement of the Utility-Privacy Trade-off

In the previous section, we present CENTAUR for the general FRL problems (3). Due to the lack of structure information, for a general non-convex optimization problem we cannot expect any utility guarantee beyond the convergence to a stationary point. In this section, we consider a specific instance of the FRL problems where both the representation function ϕ and the local classifiers h are *linear* w.r.t. their parameters $\boldsymbol{b}$ and $\boldsymbol{w}_i$. This model has been commonly used in the analysis of representation learning (Collins et al., 2021; 2022), meta learning (Tripuraneni et al., 2021; Du et al., 2020; Thekumparampil et al., 2021; Sun et al., 2021), model personalization (Jain et al., 2021), multi-task learning (Maurer et al., 2016). For this (still nonconvex) instance, we prove that CENTAUR converges to a ball centered around the *global minimizer* in a linear rate where the size of the ball depends on the required RDP parameters (α, ϵ). Let ϵ_a be the utility and let ϵ_{dp} be the DP-parameter (CENTAUR is an (ϵ_{dp}, δ)-DP mechanism). We obtain the improved privacy-utility tradeoff $\epsilon_a \cdot \epsilon_{dp} \geq O(d/n)$, which is $O(\sqrt{d})$ better than the current SOTA result, $\epsilon_a \cdot \epsilon_{dp} \geq O(d^{1.5}/n)$, by (Jain et al., 2021). In the following, we will first review objective (3) in the linear representation setting, and analyse CENTAUR to show the improved utility-privacy tradeoff.

Federated Linear Representation Learning (LRL) Recall the problem formulation in Section 3. For simplicity, we assume $m_i = m$ for all $i \in \{1, \ldots, n\}$ for some constant m. Consider the LRL setting, where given the input $\boldsymbol{x}_{ij}$ the response $y_{ij} \in \mathbb{R}$ satisfies $y_{ij} = \boldsymbol{w}_i^{*\top} \boldsymbol{B}^{*\top} \boldsymbol{x}_{ij}$. Here $\boldsymbol{B}^* \in \mathbb{R}^{d \times k}$ is a *column orthonormal* global representation matrix and $\boldsymbol{w}_i^* \in \mathbb{R}^k$ is an agent-specific optimal local linear classifiers. In terms of the notations, $\boldsymbol{B}$ corresponds to the parameter $\boldsymbol{b}$ in the general formulation (3), but is capitalized since it is now regarded as a matrix. The feasible domain $\mathbb{B}$ is the set of column orthonormal matrices $\mathbb{O}_{d,k} = \{\boldsymbol{B} \in \mathbb{R}^{d \times k} | \boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{I}_k\}$.
Given a local dataset $\mathbb{S}$ (could be $\mathbb{S}_i$ or its subset), define

$$l([\boldsymbol{w}, \boldsymbol{B}]; \mathbb{S}) = \frac{1}{|\mathbb{S}|} \sum_{(\boldsymbol{x}, y) \in \mathbb{S}} \frac{1}{2} (\boldsymbol{w}^\top \boldsymbol{B}^\top \boldsymbol{x} - y)^2. \tag{4}$$

Our goal is to recover the ground truth representation matrix $\boldsymbol{B}^*$ using the collection of the local datasets $\{\mathbb{S}_i\}_{i=1}^n$ via solving the following linear instance of the FRL problem equation 3

$$\min_{\boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{I}_k} \frac{1}{n} \sum_{i=1}^n \min_{\boldsymbol{w}_i} l([\boldsymbol{w}_i, \boldsymbol{B}]; \mathbb{S}_i) \tag{L-FRL}$$

in a federated and DP manner. Here, equation L-FRL is an instance of the general FRL problem (3) with h and ϕ set to linear functions and ℓ set to the least square loss. Note that, despite the (relatively) simple structure of equation L-FRL, it is still *non-convex* w.r.t. the variable $\boldsymbol{B}$.

Methodology for the LRL case. To establish our novel privacy and utility guarantees, we need to specify INITIALIZATION and AGGREGATION in the procedures SERVER and we also need to slightly modify the CLIENT procedure, which are elaborated as follows.

1. To obtain the novel guarantee of converging to the global minimizer for the LRL case, the initializer $\boldsymbol{B}^0$ needs to be within a constant distance to the global optimal $\boldsymbol{B}^*$ which requires a more

Algorithm 3 INITIALIZATION procedure for CENTAUR in the LRL case.

procedure INITILIZATION($T_0, \epsilon_i, L_0, \sigma_0, \zeta_0$)
  Run T_0 independent copies of PPM(L, σ_0, ζ_0) to obtain T_0 candidates $\{\boldsymbol{B}_c\}_{c=1}^{T_0}$;
  // PPM stands for private power method and is presented in Algorithm 5 in the appendix.
  // Boost this probability to $1 - n^{-k}$ via cross validation.
  Find $\hat{c}$ in $[T_0]$ such that for at least half of $c \in [T_0]$, $s_i(\boldsymbol{B}_c^\top \boldsymbol{B}_{\hat{c}}) \geq 1 - 2\epsilon_i^2, \forall i \in [k]$.
  return $\boldsymbol{B}_{\hat{c}}$.

Algorithm 4 CLIENT procedure for CENTAUR in the LRL case

1: **procedure** CLIENT($\boldsymbol{B}^t$, $\bar{m}$)
2: Sample without replacement two subsets $\mathbb{S}_i^{t,1}$ and $\mathbb{S}_i^{t,2}$ from $\mathbb{S}_i$ both with cardinality $\bar{m}$.
3: [Phase 1:] Update the local head $\boldsymbol{w}_i^{t+1} = \arg\min_{\boldsymbol{w} \in \mathbb{R}^k} l(\boldsymbol{w}, \boldsymbol{B}^t; \mathbb{S}_i^{t,1})$.
4: [Phase 2:] Compute the local gradient of the representation $\boldsymbol{G}_i^t = \partial_{\boldsymbol{B}} l([\boldsymbol{w}_i^{t+1}, \boldsymbol{B}^t]; \mathbb{S}_i^{t,2})$;
5: [Phase 3:] **return** $\boldsymbol{G}_i^t$

sophisticated procedure than the simple random initialization. We show that this requirement can be ensured by running a modified instance of the private power method (PPM) by (Hardt & Price, 2014), but the utility guarantee only holds *with a constant probability* (Lemma F.1). A key contribution of our work is to propose a novel *cross-validation* scheme to boost the success probability of the PPM with a small extra cost. Our scheme only takes as input the outputs of independent PPM trials, and hence can be treated as post-processing, which is free of privacy risk (Lemma F.3). The proposed INITIALIZATION procedure is presented in Algorithm 3. The analyses are deferred to Appendix F.

2. As discussed earlier, the feasible domain $\mathbb{B}$ is the set of column orthonormal matrices. In order to ensure the feasibility of $\boldsymbol{B}^{t+1}$, we set AGGREGATION($\boldsymbol{B}^t, \{\boldsymbol{G}_i^t\}_{i \in \mathbb{C}^t}, \eta_g$) $= \mathcal{QR}(\boldsymbol{B}^t - \eta_g \cdot \boldsymbol{G}^t)$, where $\mathcal{QR}(\cdot)$ denotes the QR decomposition and only returns the Q matrix.

3. We make a small modification in line 2 where two subsets $\mathbb{S}_i^{t,1}$ and $\mathbb{S}_i^{t,2}$ of the local dataset are sampled *without replacement* from $\mathbb{S}_i$ and are used to replace $\mathbb{S}_i$ in Phases 1 and 2. This modification is required to establish our novel utility result, which ensures that clipping threshold of the Gaussian mechanism (line 5) in the CLIENT procedure is never reached with a high probability (Lemma C.1).

5.1 ANALYSIS OF CENTAUR IN THE LRL SETTING

Use $s_i(\boldsymbol{A})$ to denote the ith largest singular value of a matrix $\boldsymbol{A}$. Let $\boldsymbol{W}^* \in \mathbb{R}^{n \times k}$ be the collection of the local optimal classifier heads with $\boldsymbol{W}_{i,:}^* = \boldsymbol{w}_i^*$. We use s_i as a shorthand for $s_i(\boldsymbol{W}^*/\sqrt{n})$ and use $\kappa = s_1/s_k$ to denote the condition number of the problem. We choose the scaling $1/\sqrt{n}$ such that s_i remains meaningful as $n \to \infty$. We make the following assumptions.

Assumption 5.1. *$\boldsymbol{x}_{ij}$ is zero mean, $\boldsymbol{I}_d$-covariance, 1-subgaussian(defined in Appendix A.1).*

Assumption 5.2. *There exists a constant $\mu > 0$ such that $\max_{i \in [n]} \|\boldsymbol{w}_i^*\|_2 \leq \mu\sqrt{k} s_k$.*

Assumption 5.3. *The number of local data points is sufficiently large: $m \geq \tilde{c}_{ld} \max(k^2, d/n)$, and the number of global data points is sufficiently large: $mn \geq \tilde{c}_{gd} k^4 d$. Here $\tilde{c}_{ld}$ and $\tilde{c}_{gd}$ hide dependence on κ, μ, and the log factors.*

These assumptions are standard in literature (Li et al., 2020b; Collins et al., 2021; Jain et al., 2021). While Problem (L-FRL) is non-convex, we show that the consensus representation $\boldsymbol{B}^t$ converges to a ball centered around the ground truth solution $\boldsymbol{B}^*$ in a linear rate. The size of the ball is controlled by the noise multiplier σ_g, which will be the free parameter that controls the utility-privacy tradeoff.

High level idea. For Problem (L-FRL), we find an initial point close to the ground truth solution via the method of moments. Given this strong initialization, CENTAUR converges linearly to the vicinity of the ground truth since it can be interpreted as an inexact gradient descent method with a fast decreasing bias term. One caveat that requires particular attention is the clipping step in the Gaussian mechanism (line 5 in Algorithm 4) will destroy the above interpretation if the threshold parameter ζ_g is set too small. To resolve this, we set ζ_g to be a high probability upper bound of $\|\boldsymbol{G}_i^t\|_F$ so that the clipping step only takes effect with a negligible probability.

In the utility analysis of the LRL case, we use the principal angle distance to measure the convergence of the column-orthonormal variable $\boldsymbol{B}$ towards the ground truth $\boldsymbol{B}^*$. We also refer to this quantity as the *utility* of the algorithm. Let $\boldsymbol{B}_\perp$ be the orthogonal complement of $\boldsymbol{B}$. We define

$$\mathrm{dist}(\boldsymbol{B}, \boldsymbol{B}^*) := \|(\boldsymbol{I}_d - \boldsymbol{B}\boldsymbol{B}^\top)\boldsymbol{B}^*\|_2 = \|\boldsymbol{B}_\perp^\top \boldsymbol{B}^*\|_2.$$

To simplify the presentation of the result, we make the following assumptions: The dimension of the input is sufficiently large $d \geq \kappa^8 k^3 \log n$ and the number of clients is sufficiently large $n \geq m$. The proof of the following theorem can be find in Appendix C.

Theorem 5.1 (Utility Analysis)**.** *Consider the instance of* CENTAUR *for the LRL setting with its* CLIENT *and* INITIALIZATION *procedures defined in Algorithms 4 and 3 respectively. Suppose that the matrix $\boldsymbol{B}^0$ returned by* INITIALIZATION *satisfies $\mathrm{dist}(\boldsymbol{B}^0, \boldsymbol{B}^*) \leq \epsilon_0 = 0.2$, and suppose that the mini-batch size parameter $\bar{m}$ satisfies $\bar{m} \geq c_m \max\{\kappa^2 k^2 \log n, k^2 d/n\}$. Set the clipping threshold of the Gaussian mechanism $\zeta_g = c_\zeta \mu^2 k s_k^2 \sqrt{dk \log n}$, the global step size $\eta_g = 1/4s_1^2$, the number of global rounds $T_g = c_T \kappa^2 \log(\kappa \eta_g \zeta_g \sigma_g d/n)$. Assuming that the noise multiplier in the Gaussian mechanism is sufficiently small: $\sigma_g \leq c_\sigma n \kappa^4/(\mu^2 d\sqrt{k \log n})$. We have with probability at least $1 - c_p \bar{m} T_g \cdot n^{-k}$,*

$$\mathrm{dist}(\boldsymbol{B}^{T_g}, \boldsymbol{B}^*) \leq c_d \kappa^2 \eta_g \sigma_g \zeta_g \sqrt{d}/n = \tilde{c}_d \sigma_g \mu^2 k^{1.5} d/n. \tag{5}$$

Here c_m, c_ζ, c_T, c_σ, c_p and c_d are universal constants.

As will be discussed in Theorem 5.2, a larger noise multiplier σ_g leads to a smaller privacy risk. On the other hand, in order for the above utility guarantee to hold, we cannot have arbitrarily good privacy protection as that would violate the limitation on σ_g. The intuition behind this requirement is that our convergence analysis requires the iterates to stay within a ball centered around the ground truth, with a constant radius (measured in terms of the principal angle distance). Adding a large Gaussian noise will break this argument. Similar requirements are made in Tripuraneni et al. (2021).

Since the SERVER procedure remains exactly the same as Algorithm 1 in the LRL case, the main body (anything after INITIALIZATION) of the resulting CENTAUR instance has the same privacy guarantee as described in remark 4.1. However, we still need to account for the privacy leakage of the INITIALIZATION procedure in Algorithm 3 as it is data-dependent. This will be deferred to Appendix F, where we show that Algorithm 3 is an $(\alpha, \epsilon_{init})$-RDP mechanism, with ϵ_{init} defined in Corollary F.1. Combining this fact with the RDP analysis for the main body leads to the following privacy guarantee for CENTAUR in the LRL case (see Appendix A.4).

Theorem 5.2 (Privacy Bound)**.** *Consider the instance of* CENTAUR *with its* CLIENT *procedure defined in Algorithm 4 and its* INITIALIZATION *procedure defined in Algorithm 3. Suppose that the* INITIALIZATION *procedure is an $(\alpha, \epsilon_{init})$-RDP mechanism (proved in Corollary F.1). Let σ_g, the noise multiple in the* CLIENT *procedure, be a free parameter that controls the privacy-utility trade-off. By setting the inputs to* SERVER *as $p_g = 1$, $T_g = c_T \kappa^2 \log(\kappa \eta_g \zeta_g \sigma_g d/n)$, the instance of* CENTAUR *under consideration is an $(\alpha, \epsilon_{init} + \epsilon_{rdp}/2)$-RDP mechanism, where $\epsilon_{rdp} = 4\alpha T_g/\sigma_g^2$. Moreover, when $\sigma_g = \tilde{O}(n/(k^4\mu^2 d))$, we have $\epsilon_{init} \leq \epsilon_{rdp}/2$, in which case* CENTAUR *is an (α, ϵ_{rdp})-RDP mechanism and is also an (ϵ_{dp}, δ)-DP mechanism with $\epsilon_{dp} = 2\sqrt{T_g \log(1/\delta)}/\sigma_g$.*

5.2 OVERALL UTILITY-PRIVACY TRADE-OFF

We now combine the utility and privacy analyses of CENTAUR in the LRL setting to obtain the overall utility-privacy trade-off in the following sense: According to Theorem 5.1, to achieve a high utility, i.e. a small ϵ_a, we need to choose a small noise multiplier σ_g while Theorem 5.2 states that the smaller σ_g is, the larger the privacy cost. For the simplicity of the presentation, we focus on the regime where the trade-off parameter satisfies $\sigma_g = \tilde{O}(n/k^4\mu^2 d)$.

Corollary 5.1. *Use ϵ_a to denote the target utility, i.e. $\mathrm{dist}(\boldsymbol{B}^T, \boldsymbol{B}^*) \leq \epsilon_a$ where $\boldsymbol{B}^T$ is the output of* CENTAUR *and use ϵ_{dp} to denote the privacy budget, i.e.* CENTAUR *is an (ϵ_{dp}, δ)-DP mechanism. We have the following tradeoff where $\tilde{c}_t$ hides the constants and* log *terms:*

$$\tilde{c}_t \kappa k^{1.5} \mu^2 d/n \leq \epsilon_a \cdot \epsilon_{dp}.$$

When focusing on the input dimension d and the number of clients n and treating other factors as constants, the above trade-off can be simplified to $O(d/n) \leq \epsilon_a \cdot \epsilon_{dp}$. Recall that the previous SOTA result of the LRL setting by Jain et al. (2021) (Lemma 4.4 therein) can be interpreted as $O(d^{1.5}/n) \leq \epsilon_a \cdot \epsilon_{dp}$. Hence, we obtain an $O(\sqrt{d})$ improvement.

Table 1: Testing accuracy (%) on CIFAR10/CIFAR100/EMNIST with various data allocation settings. No data augmentation is used. n stands for the number clients and S stands for the number classes per client. The δ parameter of DP is fixed to 10^{-5} as a common choice in the literature. The budget parameter of DP is fixed to a small value of 1 for results in this table, i.e. $\epsilon_{dp} = 1$.

Methods	`Stand-alone-no-FL`	`DP-FedAvg-ft`	`PMTL-ft`	`PPSGD`	`CENTAUR`
CIFAR10 n=500, S=5	50.03 (1.59)	38.04 (1.97)	42.06 (0.94)	47.94 (1.87)	**52.55 (0.69)**
CIFAR10 n=1000,S=2	74.06 (0.45)	57.02 (1.36)	67.79 (1.77)	71.52 (2.30)	**75.75 (0.20)**
CIFAR10 n=1000,S=5	44.60 (1.30)	37.15 (1.22)	35.99 (1.79)	44.61 (2.68)	**47.67 (1.62)**
CIFAR100 n=1000,S=5	39.20 (1.12)	22.17 (2.12)	24.17 (1.43)	38.40 (0.66)	**41.59 (1.17)**
EMNIST n=1000,S=5	93.47 (0.14)	91.32 (0.22)	92.32 (0.22)	93.44 (0.20)	**94.17 (0.19)**
EMNIST n=2000,S=5	90.67 (0.46)	86.85 (1.11)	88.81 (2.08)	88.96 (1.93)	**92.79 (0.25)**

6 EXPERIMENTS

In this section, we present the empirical results that show the significant advantage of the proposed `CENTAUR` over previous arts. Four baselines are included: `Stand-alone-no-FL` which stands for local stand-alone training; `DP-FedAvg-ft` which stands for DP-FedAvg with local fine tuning (Yu et al., 2020); `PPSGD` proposed by Bietti et al. (2022); and `PMTL-ft` which stands for PMTL proposed by Hu et al. (2021) with local fine tuning. Note that `Stand-alone-no-FL` does not involve any global communications, therefore no privacy mechanism is added to its implementation. This makes `Stand-alone-no-FL` a strong baseline as the utility of all included differentially private competing methods are affected by gradient clipping and noise injection, especially when the DP budget is small, e.g. $\epsilon_{dp} = 1$. Another advantage of `Stand-alone-no-FL` setting is that, the local stand-alone models are highly flexible, i.e. the model on one client and be completely different from the one on others. On the contrary, while the models of all other non-local methods share a common representation part, which takes up the major portion of the whole model.

We focus on the task of image classification and conduct experiments on three representative datasets, namely CIFAR10, CIFAR100, and EMNIST. In terms of architecture of the neural network, we use LeNet for CIFAR10/CIFAR100 and use MLP for EMNIST, as commonly used in the federated learning literature, the details of which are discussed in Appendix. In terms of data augmentation, we do not perform any data augmentation, as we observe that naive data augmentation for DP training leads to worse accuracy, as also reported in De et al. (2022).

`CENTAUR` Has the Best Privacy Utility Trade-off. We first present the utility (testing accuracy) of models trained with `CENTAUR` and other baselines algorithms under a fixed small privacy budget $\epsilon_{dp} = 1$, for a variety of heterogeneous FL settings. To simulate the data-heterogeneity phenomenon ubiquitous in the research of federated learning, we follow the data allocation scheme of (Collins et al., 2021): Specifically, we first split the original dataset into a training part (90%) and a validation part (10%) and we then allocate the training part equally to n clients while ensuring that each client has at most data from S classes. In Table 1, we observe that, under this small privacy budget, our proposed `CENTAUR` enjoy better performance than all the included baseline algorithms. Importantly, `CENTAUR` is **the only method** that consistently **outperforms the strong local-only baseline**, and therefore justifies the choice of collaborative learning as opposed of local stand-alone training. Finally, we further demonstrate that `CENTAUR` enables superior privacy utility trade-off uniformly across different privacy budget ϵ, for the setting of EMNIST dataset ($n = 2000, S = 5$) in Figure 1.

Conclusion. In this work, we point out that the phenomenon of statistical heterogeneity, one of the major challenges of federated learning, introduces extra difficulty when DP constraints are imposed. To alleviate this difficulty, we consider the federated representation learning where only the representation function is to be globally shared and trained. We provide a rigorous guarantee for the utility-privacy trade-off of the proposed `CENTAUR` method in the linear representation setting, which is $O(\sqrt{d})$ better than the SOTA result. We also empirically show that `CENTAUR` provides better utility uniformly on several vision datasets under various data heterogeneous settings.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2020.

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.

Naman Agarwal, Ananda Theertha Suresh, Felix Xinnan X Yu, Sanjiv Kumar, and Brendan McMahan. cpsgd: Communication-efficient and differentially-private distributed sgd. *Advances in Neural Information Processing Systems*, 31, 2018.

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Alberto Bietti, Chen-Yu Wei, Miroslav Dudik, John Langford, and Steven Wu. Personalization improves privacy-accuracy tradeoffs in federated learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1945–1962. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/bietti22a.html`.

Zachary Charles and Jakub Konečnỳ. Convergence and accuracy trade-offs in federated learning and meta-learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2575–2583. PMLR, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pp. 2089–2099. PMLR, 2021.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Fedavg with fine tuning: Local updates lead to representation learning. *NeurIPS*, 2022.

Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. *Advances in Neural Information Processing Systems*, 30, 2017.

Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Privacy aware learning. *Journal of the ACM (JACM)*, 61(6):1–57, 2014.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pp. 1054–1067, 2014.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020.

Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. *Advances in neural information processing systems*, 27, 2014.

Justin Hsu, Zhiyi Huang, Aaron Roth, Tim Roughgarden, and Zhiwei Steven Wu. Private matchings and allocations. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '14, pp. 21–30, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450327107. doi: 10.1145/2591796.2591826. URL https://doi.org/10.1145/2591796.2591826.

Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. Private multi-task learning: Formulation and applications to federated learning. *arXiv preprint arXiv:2108.12978*, 2021.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 665–674, 2013.

Prateek Jain, John Rush, Adam Smith, Shuang Song, and Abhradeep Guha Thakurta. Differentially private model personalization. *Advances in Neural Information Processing Systems*, 34, 2021.

Yihan Jiang, Jakub Konečnỳ, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019.

Peter Kairouz, Brendan McMahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu. Practical and private (deep) learning without sampling or shuffling. In *International Conference on Machine Learning*, pp. 5213–5225. PMLR, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Daniel Levy, Ziteng Sun, Kareem Amin, Satyen Kale, Alex Kulesza, Mehryar Mohri, and Ananda Theertha Suresh. Learning with user-level privacy. *Advances in Neural Information Processing Systems*, 34:12466–12479, 2021.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020b.

Zhu Li, Zhi-Hua Zhou, and Arthur Gretton. Towards an understanding of benign overfitting in neural networks. *arXiv preprint arXiv:2106.03212*, 2021.

Yingyu Liang, Maria-Florina F Balcan, Vandana Kanchanapally, and David Woodruff. Improved distributed principal component analysis. *Advances in neural information processing systems*, 27, 2014.

Grigory Malinovskiy, Dmitry Kovalev, Elnur Gasanov, Laurent Condat, and Peter Richtarik. From local sgd to local fixed-point methods for federated learning. In *International Conference on Machine Learning*, pp. 6692–6701. PMLR, 2020.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017a.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017b.

Ilya Mironov, Kunal Talwar, and Li Zhang. R\'enyi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

Aritra Mitra, Rayana Jaafar, George J Pappas, and Hamed Hassani. Achieving linear convergence in federated learning under objective and systems heterogeneity. *NeurIPS*, 2021.

Milad Nasr, Reza Shokri, and Amir Houmansadr. 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)*, pp. 739–753. IEEE, 2019.

Krishna Pillutla, Kshitiz Malik, Abdel-Rahman Mohamed, Mike Rabbat, Maziar Sanjabi, and Lin Xiao. Federated learning with partial model personalization. In *International Conference on Machine Learning*, pp. 17716–17758. PMLR, 2022.

Mark Rudelson and Roman Vershynin. The littlewood–offord problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600–633, 2008.

Zebang Shen, Hamed Hassani, Satyen Kale, and Amin Karbasi. Federated functional gradient boosting. In *International Conference on Artificial Intelligence and Statistics*, pp. 7814–7840. PMLR, 2022.

Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1310–1321, 2015.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. *Advances in neural information processing systems*, 30, 2017.

Yue Sun, Adhyyan Narang, Ibrahim Gulluk, Samet Oymak, and Maryam Fazel. Towards sample-efficient overparameterized meta-learning. *Advances in Neural Information Processing Systems*, 34:28156–28168, 2021.

Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Statistically and computationally efficient linear meta-representation learning. *Advances in Neural Information Processing Systems*, 34:18487–18500, 2021.

Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pp. 10434–10443. PMLR, 2021.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Kangkang Wang, Rajiv Mathews, Chloé Kiddon, Hubert Eichner, Françoise Beaufays, and Daniel Ramage. Federated evaluation of on-device personalization. *arXiv preprint arXiv:1910.10252*, 2019.

Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. Salvaging federated learning by local adaptation. *arXiv preprint arXiv:2002.04758*, 2020.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.