DePRL: Achieving Linear Convergence Speedup in Personalized Decentralized Learning with Shared Representations

Guojun Xiong[1], **Gang Yan**[2], **Shiqiang Wang**[3], **Jian Li**[1]
[1]Stony Brook University
[2]Binghamton University
[3]IBM T. J. Watson Research Center
{guojun.xiong, jian.li.3}@stonybrook.edu, gyan2@binghamton.edu, wangshiq@us.ibm.com

Abstract

Decentralized learning has emerged as an alternative method to the popular parameter-server framework which suffers from high communication burden, single-point failure and scalability issues due to the need of a central server. However, most existing works focus on a single shared model for all workers regardless of the data heterogeneity problem, rendering the resulting model performing poorly on individual workers. In this work, we propose a novel personalized decentralized learning algorithm named DePRL via shared representations. Our algorithm relies on ideas from representation learning theory to learn a low-dimensional global representation collaboratively among all workers in a fully decentralized manner, and a user-specific low-dimensional local head leading to a personalized solution for each worker. We show that DePRL achieves, for the first time, a provable *linear speedup for convergence* with general non-linear representations (i.e., the convergence rate is improved linearly with respect to the number of workers). Experimental results support our theoretical findings showing the superiority of our method in data heterogeneous environments.

Introduction

Fueled by the rise of machine learning applications in Internet of Things, federated learning (FL) (McMahan et al. 2017; Imteaj et al. 2022) has become an emerging paradigm that allows a large number of workers to produce a global model without sharing local data. The task of coordinating between workers is fulfilled by a central server that aggregates models received from workers at each round and broadcasts updated models to them. However, this parameter-server (PS) based scheme has a major drawback for the need of a central server (Kairouz et al. 2019). In practice, the communication occurs between the server and workers leads to a quite large communication burden for the server (Lian et al. 2017), and the server could face system failure or attacks, which may leak users' privacy or jeopardize the training process.

With this regard, *consensus-based decentralized learning* has recently emerged as a promising method, where each worker maintains a local copy of the model and embraces peer-to-peer communication for faster convergence (Lian et al. 2017, 2018). In decentralized learning, workers follow a communication graph to reach a so-called consensus model. However, like conventional PS framework, one of the most important challenges in decentralized learning is the issue of *data heterogeneity*, where the data distribution among workers may vary to a large extent. As a result, if all workers learn a *single shared model* with parameter $\mathbf{w}$, the resulting model could perform poorly on many of individual workers. To this end, *personalized decentralized learning* (Vanhaesebrouck, Bellet, and Tommasi 2017; Dai et al. 2022) is important for achieving personalized models for each worker i with parameter $\mathbf{w}_i$ instead of using a single shared model.

In this paper, we take a further step towards personalized decentralized learning. In particular, we take advantage of common representation among workers. This is inspired by observations in centralized learning, which suggest that heterogeneous data distributed across tasks (e.g., image classification) may share a common (low-dimensional) representation despite having different labels (Bengio, Courville, and Vincent 2013; LeCun, Bengio, and Hinton 2015). To our best knowledge, Collins et al. (2021) is the first to leverage this insight to design personalized PS based scheme, while we generalize it to decentralized setting. Specifically, we consider the setting in which all workers' model parameters share a common map, coupled with a personalized map that fits their local data. Formally, the parameter for worker i's model can be represented as $\mathbf{w}_i = \boldsymbol{\theta}_i \circ \boldsymbol{\phi}$, where $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^z$ is a shared **global representation**[1] which maps d-dimensional data points to a lower space of size z, and $\boldsymbol{\theta}_i : \mathbb{R}^z \to \mathcal{Y}$ is the worker specific **local head** which maps from the lower dimensional subspace to the space of labels. Typically $z \ll d$ and thus given any fixed representation $\boldsymbol{\phi}$, the worker specific heads $\boldsymbol{\theta}_i$ are easy to optimize locally. Though Collins et al. (2021) provided a rigorous analysis with linear global representation, the following important questions remain open:

Does there exist a personalized, fully decentralized algorithm that can solve the optimization problem $\min_{\boldsymbol{\phi}\in\Phi} \frac{1}{N}\sum_{i=1}^{N} \min_{\boldsymbol{\theta}_i\in\Theta} F_i(\boldsymbol{\theta}_i \circ \boldsymbol{\phi})$*, where* $F_i(\cdot)$ *is the loss function associated with worker* i*? Can we provide a convergence analysis for such a personalized, decentralized algo-*

[1]For abuse of notion, we use $\boldsymbol{\phi}$ to denote both the global representation model and its associated parameter, and $\{\boldsymbol{\theta}_i\}_{i=1}^{N}$ to denote both the local heads and its associated parameter. For simplicity, we call $\boldsymbol{\phi}$ the *global representation* and $\boldsymbol{\theta}_i$ the *local head* of worker i in the rest of the paper. The "$\circ$" symbol denotes the composition relation between the parameters $\boldsymbol{\theta}_i$ and $\boldsymbol{\phi}$ as in Collins et al. (2021).

rithm under general non-linear representations?

In this paper, we provide affirmative answers to these questions. We propose *a fully decentralized algorithm* named DePRL with alternating updates between global representation and local head parameters to solve the above optimization. At each round, each worker performs one or more steps of stochastic gradient descent to update its local head and global representation from its side. Then each worker *only* shares its updated global representation with a subset of workers (neighbors) in the communication graph and computes a weighted average (i.e., consensus component) of global representations received from its neighbors. The updated local head and global representation after consensus serve as the initialization for the next round update. All workers in DePRL collaborate to learn a common global representation, while locally each worker learns its unique head.

Compared to conventional decentralized learning with a single shared model (Lian et al. 2017, 2018; Assran et al. 2019), the updates of parameters of local head and global representation in DePRL are strongly coupled due to their intrinsic dependence and iterative update nature. This makes existing convergence analysis for decentralized learning with a single shared model not directly applicable to ours, and necessitates different proof techniques. One fundamental reason is that, instead of learning only a single shared model, there are multiple local heads that need to be handled in DePRL and the updates of local heads are also strongly coupled with global representation. We summarize our contributions:

• **DePRL Algorithm.** We propose for the first time *a fully decentralized* algorithm named DePRL which leverages ideas from representation learning theory to learn a global representation collaboratively among all workers, and a user-specific local head leading to a personalized solution for each worker.

• **Convergence Rate.** To incorporate the impact of two coupled parameters, we first introduce a new notion of ϵ-approximation solution. Using this notion, to our best knowledge, we provide the first convergence analysis of personalized decentralized learning with *shared non-linear representations*. We show that the convergence rate of DePRL is $\mathcal{O}(\frac{1}{\sqrt{NK}})$, where N is the number of workers, and K is the number of communication rounds which is sufficiently large. This indicates that DePRL achieves a **linear speedup** for convergence with respect to the number of workers. This is the first linear speedup result for personalized decentralized learning with shared representations, and is highly desirable since it implies that one can efficiently leverage the massive parallelism in large-scale decentralized systems. In addition, interestingly, our results guarantee that all workers reach a consensus on the shared global representation, while learn a personalized local head. This reveals new insights on the relationship between personalized decentralized model with shared representations and its generalization to unseen workers that have not participated in the training process, as we numerically verify in experimental results.

• **Evaluation.** To examine the performance of DePRL and verify our theoretical results, we conduct experiments on different datasets with representative DNN models and compare with a set of baselines. Our results show the superior performance of DePRL in data heterogeneous environments.

System Model and Problem Formulation

Notation. Denote the number of workers and communication rounds as N and K, respectively. We use calligraphy letter $\mathcal{A}$ to denote a finite set with cardinality $|\mathcal{A}|$, and $[N]$ to denote the set of integers $\{1, \cdots, N\}$. We use boldface to denote matrices and vectors, and $\|\cdot\|$ to denote the l_2-norm.

Consensus-based Decentralized Learning. Supervised learning aims to learn a model with optimal parameter that maps an input to an output by using examples from a training data set $\mathcal{D}$ with each example being a pair of input $\mathbf{x}_m$ and the associated output y_m. Due to increases in available data and the complexity of statistical model, an efficient decentralized algorithm is to offload the computation overhead to N workers, which jointly determine the optimal parameters through a decentralized coordination. This gives rise to the minimization of the sum of functions local to each worker

$$\min_{\mathbf{w}} f(\mathbf{w}) := \frac{1}{N}\sum_{i=1}^{N} F_i(\mathbf{w}), \tag{1}$$

where $F_i(\mathbf{w}) = \frac{1}{|\mathcal{D}_i|}\sum_{(\mathbf{x}_m, y_m)\in\mathcal{D}_i} \ell(\mathbf{w}, \mathbf{x}_m, y_m)$, $\mathcal{D}_i$ is worker i's local dataset, with $\ell(\mathbf{w}, \mathbf{x}_m, y_m)$ being model error on example $(\mathbf{x}_m, y_m)$ using model parameter $\mathbf{w} \in \mathbb{R}^{d\times 1}$. The decentralized system can be modeled as *a communication graph* $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ with $\mathcal{N} = [N]$ being the set of workers and an edge $(i, j) \in \mathcal{E}$ indicates that workers i and j can communicate with each other. We assume the graph is strongly connected (Nedic and Ozdaglar 2009; Nedić, Olshevsky, and Rabbat 2018), i.e., there exists at least one path between any two arbitrary workers. Denote neighbors of worker i as $\mathcal{N}_i = \{j|(j, i) \in \mathcal{E}\} \cup \{i\}$. All workers perform local updates synchronously and broadcast updated models to their neighbors. Each worker then computes a weight average (i.e., consensus component) of the received models from its neighbors, which serves as the initialization for next round.

Personalization via Common Representation. Conventional decentralized learning aims at learning a *single* shared model parameter $\mathbf{w}$ that performs well on average across all workers (Lian et al. 2017, 2018). However, this approach may yield a solution that performs poorly in heterogeneous settings where data distributions vary across workers. Indeed, in the presence of data heterogeneity, the error functions F_i will have different minimizers. This necessities the search for more personalized solutions $\{\mathbf{w}_i\}_{i=1}^N$ that can be learned in a decentralized manner using workers' local data.

To address this challenge, Collins et al. (2021) leveraged representation learning theory into the PS setting. Formally, the parameter for worker i's model can be represented as $\mathbf{w}_i = \boldsymbol{\theta}_i \circ \boldsymbol{\phi}$, where $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^z$ is a shared ***global representation*** which maps d-dimensional data points to a lower space of size z, and $\boldsymbol{\theta}_i : \mathbb{R}^z \to \mathcal{Y}$ is the worker specific ***local head*** which maps from the lower dimensional subspace to the space of labels. See an illustrative example of Collins et al. (2021) in Xiong et al. (2023a). We generalize this common structure studied in Collins et al. (2021) to the decentralized setting, with which, (1) can be reformulated as

$$\min_{\boldsymbol{\phi}\in\Phi}\min_{\boldsymbol{\theta}_i\in\Theta} f(\boldsymbol{\phi}, \{\boldsymbol{\theta}_i\}_{i=1}^N) := \frac{1}{N}\sum_{i=1}^{N} F_i(\boldsymbol{\theta}_i \circ \boldsymbol{\phi}), \tag{2}$$

where Φ is the class of feasible representations and Θ is the class of feasible heads. In our proposed decentralized learning scheme, workers collaborate to learn global representation $\boldsymbol{\phi}$ using all workers' data, while using their local information to learn personalized local heads $\{\boldsymbol{\theta}_i\}_{i=1}^N$. In other words, worker i maintains *a local estimate* of global representation $\boldsymbol{\phi}_i(k)$ at each round k and broadcasts it to its neighbors in $\mathcal{N}_i$, while the local head $\boldsymbol{\theta}_i(k)$ is only updated locally.

DePRL Algorithm

In personalized decentralized learning, workers aim to learn the global representation $\boldsymbol{\phi}$ collaboratively, while each worker i aims to learn a unique local head $\boldsymbol{\theta}_i$ locally. To achieve this, we propose a stochastic gradient descent (SGD)-based algorithm named DePRL that solves (2) ***in a fully decentralized manner***. Specially, DePRL alternates between three steps among all workers at each communication round: (a) local head update; (b) local representation update; and (c) consensus-based global representation update.

Local Head Update. At round k, worker i makes τ local stochastic gradient-based updates to solve for its optimal local head $\boldsymbol{\theta}_i(k)$ given the current global representation $\boldsymbol{\phi}_i(k)$ on its local side. In other words, for $s = 0, \ldots, \tau-1$, worker i updates its local head as

$$\boldsymbol{\theta}_i(k, s+1) = \boldsymbol{\theta}_i(k, s) - \alpha g_{\boldsymbol{\theta}}(\boldsymbol{\phi}_i(k), \boldsymbol{\theta}_i(k, s)), \quad (3)$$

where α is the learning rate for local head and $g_{\boldsymbol{\theta}}(\boldsymbol{\phi}_i(k), \boldsymbol{\theta}_i(k, s))$ is a stochastic gradient of local head $\boldsymbol{\theta}_i(k, s)$ given the global representation $\boldsymbol{\phi}_i(k)$ on its side:

$$g_{\boldsymbol{\theta}}(\boldsymbol{\phi}_i(k), \boldsymbol{\theta}_i(k, s)) := \frac{1}{|\mathcal{C}_i(k, s)|} \sum_{(\mathbf{x}_m, y_m) \in \mathcal{C}_i(k,s)} \nabla_{\boldsymbol{\theta}} F_i(\boldsymbol{\phi}_i(k), \boldsymbol{\theta}_i(k, s), \mathbf{x}_m, y_m), \quad (4)$$

where $\mathcal{C}_i(k, s)$ is a random subset of $\mathcal{D}_i$. We allow each worker to perform τ steps local updates to find the optimal local head based on its local data. For ease of presentation, we denote $\boldsymbol{\theta}_i(k+1) := \boldsymbol{\theta}_i(k+1, 0) = \boldsymbol{\theta}_i(k, \tau-1)$.

Local Representation Update. Once the updated local heads $\boldsymbol{\theta}_i(k+1)$ are obtained, each worker i executes one-step local update on their representation parameters, i.e.,

$$\boldsymbol{\phi}_i(k+1/2) = \boldsymbol{\phi}_i(k) - \beta g_{\boldsymbol{\phi}}(\boldsymbol{\phi}_i(k), \boldsymbol{\theta}_i(k+1)), \quad (5)$$

where β is the learning rate for global representation and $g_{\boldsymbol{\phi}}(\boldsymbol{\phi}_i(k), \boldsymbol{\theta}_i(k+1))$ is the stochastic gradient of global representation $\boldsymbol{\phi}_i(k)$ given the updated local head $\boldsymbol{\theta}_i(k+1)$:

$$g_{\boldsymbol{\phi}}(\boldsymbol{\phi}_i(k), \boldsymbol{\theta}_i(k+1)) := \frac{1}{|\mathcal{C}_i(k)|} \sum_{(\mathbf{x}_m, y_m) \in \mathcal{C}_i(k)} \nabla_{\boldsymbol{\phi}} F_i(\boldsymbol{\phi}_i(k), \boldsymbol{\theta}_i(k+1), \mathbf{x}_m, y_m). \quad (6)$$

Consensus-based Global Representation Update. Each worker i broadcasts its local representation update $\boldsymbol{\phi}_i(k+1/2)$ to its neighbors $j \in \mathcal{N}_i$, and computes a weighted average (i.e., consensus component) of local representation updates $\boldsymbol{\phi}_j(k+1/2)$ received from its neighbors j to produce the next representation model $\boldsymbol{\phi}_i(k+1)$ on its side:

$$\boldsymbol{\phi}_i(k+1) = \sum\nolimits_{j \in \mathcal{N}_i} \boldsymbol{\phi}_j(k+1/2) P_{i,j}, \quad (7)$$

Algorithm 1: DePRL

Parameters: Learning rates α, β; update step number for local head τ; number of communication rounds K.
Initialize $\boldsymbol{\phi}(0), \boldsymbol{\theta}_1(0,0), \ldots, \boldsymbol{\theta}_N(0,0)$.
for $k = 0, 1, \ldots, K-1$ **do**
  for $i = 1, \cdots, N$ **do**
    for $s = 0, \ldots, \tau-1$ **do**
      $\boldsymbol{\theta}_i(k, s+1) \leftarrow \boldsymbol{\theta}_i(k, s) - \alpha g_{\boldsymbol{\theta}}(\boldsymbol{\phi}_i(k), \boldsymbol{\theta}_i(k, s))$;
    end for
    $\boldsymbol{\phi}_i(k+1/2) \leftarrow \boldsymbol{\phi}_i(k) - \beta g_{\boldsymbol{\phi}}(\boldsymbol{\phi}_i(k), \boldsymbol{\theta}_i(k+1))$;
    $\boldsymbol{\phi}_i(k+1) \leftarrow \sum_{j \in \mathcal{N}_i} \boldsymbol{\phi}_j(k+1/2) P_{i,j}$;
    Worker i initializes $\boldsymbol{\theta}_i(k+1, 0) \leftarrow \boldsymbol{\theta}_i(k, \tau-1)$.
  end for
end for

where $\mathbf{P} = (P_{i,j})$ is a $N \times N$ non-negative *consensus matrix*.

DePRL alternates between (3), (5) and (7) at each round, and the entire procedure is summarized in Algorithm 1. An example of DePRL with 3 workers is illustrated in Figure 1.

Remark 1. *We update parameters using SGD in (4) and (6); however, DePRL can be easily incorporated with other methods such as gradient descent with momentum. Further, we perform one-step update on global representation in (5) given that we perform τ-step updates on local head in (3). However, DePRL can be easily generalized to multi-step representation update in (5). Finally, we note that our representation model is inspired by Collins et al. (2021), which considered the PS setting. The theoretical analysis in Collins et al. (2021) focused on showing that the learned representation converges to a ground-truth under the assumption that the global representation must be linear. In contrast, we consider a fully decentralized framework and our convergence analysis for DePRL is under the general non-linear representations. In addition, we numerically evaluate the generalization performance of DePRL under non-linear representations.*

Convergence Analysis

In this section, we provide a rigorous analysis of the convergence of DePRL in the decentralized setting with the general non-linear representation model.

ϵ-Approximation Solution

We first introduce the notion of ϵ-approximation solution. We denote $\bar{\boldsymbol{\phi}}(k) := \frac{1}{N}\sum_{i=1}^N \boldsymbol{\phi}_i(k)$ as the consensus global representation, and the partial gradients of the global loss function with respect to (w.r.t.) $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ as $\nabla_{\boldsymbol{\theta}} f(\bar{\boldsymbol{\phi}}(k), \{\boldsymbol{\theta}_i(k)\}_{i=1}^N)$ and $\nabla_{\boldsymbol{\phi}} f(\bar{\boldsymbol{\phi}}(k), \{\boldsymbol{\theta}_i(k+1)\}_{i=1}^N)$, respectively, satisfying

$$\nabla_{\boldsymbol{\theta}} f(\bar{\boldsymbol{\phi}}(k), \{\boldsymbol{\theta}_i(k)\}_{i=1}^N) := \frac{1}{N}\sum_{i=1}^N \nabla_{\boldsymbol{\theta}} F_i(\bar{\boldsymbol{\phi}}(k), \boldsymbol{\theta}_i(k)),$$
$$\nabla_{\boldsymbol{\phi}} f(\bar{\boldsymbol{\phi}}(k), \{\boldsymbol{\theta}_i(k+1)\}_{i=1}^N) := \frac{1}{N}\sum_{i=1}^N \nabla_{\boldsymbol{\phi}} F_i(\bar{\boldsymbol{\phi}}(k), \boldsymbol{\theta}_i(k+1)), \quad (8)$$

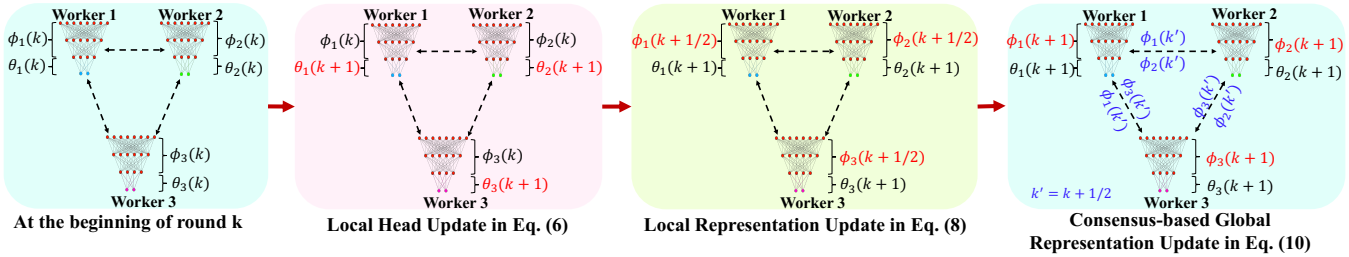

Figure 1: An illustrative example of DePRL for 3 workers with the communication graph being a ring (indicated by black dashed lines). (a) At the beginning of each round k, each worker $i = 1, 2, 3$ has the local head $\boldsymbol{\theta}_i(k)$ and the global representation $\boldsymbol{\phi}(k)$ on its side, which we denote as $\boldsymbol{\phi}_i(k)$. (b) *Local Head Update:* With $(\boldsymbol{\theta}_i(k), \boldsymbol{\phi}_i(k))$, each worker i performs τ steps SGD to obtain $\boldsymbol{\theta}_i(k+1)$. Note that $\boldsymbol{\phi}_i(k)$ remains unchanged at this step and the updated $\boldsymbol{\theta}_i(k+1)$ depends on both $\boldsymbol{\theta}_i(k)$ and $\boldsymbol{\phi}_i(k)$. (c) *Local Representation Update:* Each worker i then updates the global representation on its side by executing one-step SGD to obtain $\boldsymbol{\phi}_i(k+1/2)$, which depends on both $\boldsymbol{\theta}_i(k+1)$ and $\boldsymbol{\phi}_i(k)$. (d) *Consensus-based Global Representation Update:* Each worker i shares $\boldsymbol{\phi}_i(k+1/2)$ with its neighbors and then executes a consensus step to produce the next global representation model $\boldsymbol{\phi}_i(k+1)$. We highlight the updated parameters in each step in red, and the shared parameters (only the global representation) between workers in blue.

where we remark that in DePRL each worker i first updates its local head $\boldsymbol{\theta}_i(k)$ to $\boldsymbol{\theta}_i(k+1)$, and then updates global representation $\bar{\boldsymbol{\phi}}(k)$ provided $\boldsymbol{\theta}_i(k+1)$, see Algorithm 1.

Then, we say that $\{\{\boldsymbol{\phi}_i(k)\}_{i=1}^N, \{\boldsymbol{\theta}_i(k)\}_{i=1}^N, \forall k\}$ is an ϵ-approximation solution to (2) if it satisfies

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[M(k)] \leq \epsilon, \tag{9}$$

where

$$M(k) := \underbrace{\left\|\nabla_{\boldsymbol{\phi}} f(\bar{\boldsymbol{\phi}}(k), \{\boldsymbol{\theta}_i(k+1)\}_{i=1}^N)\right\|^2}_{\text{partial gradient w.r.t. } \boldsymbol{\phi} \text{ of global loss function}} + \frac{\alpha\tau}{\beta} \underbrace{\left\|\nabla_{\boldsymbol{\theta}} f(\bar{\boldsymbol{\phi}}(k), \{\boldsymbol{\theta}_i(k)\}_{i=1}^N)\right\|^2}_{\text{partial gradient w.r.t. } \boldsymbol{\theta} \text{ of global loss function}} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{\phi}_i(k) - \bar{\boldsymbol{\phi}}(k)\|^2}_{\text{consensus error of global representation } \boldsymbol{\phi}}. \tag{10}$$

The first two terms in (10) characterizes the performance of DePRL and the third term measures the average error of global representation from the perspective of each worker i's local representation update $\boldsymbol{\phi}_i(k)$. Since DePRL iteratively updates the local head $\{\boldsymbol{\theta}_i(k)\}_{i=1}^N$ and representation $\{\boldsymbol{\phi}_i(k)\}_{i=1}^N$ using different learning rates, i.e., τ-step updates with a rate α and one-step update with a rate β, as shown in (3) and (5), we consider weighted partial gradients w.r.t. the global loss function in the first two terms. This is inspired by finite-time analysis of two-timescale stochastic approximation (Borkar 2009). Finally, (10) does not explicitly include the local head error due to two reasons. First, we consider general non-convex loss functions, and the local optimum $\{\boldsymbol{\theta}_i^*\}_{i=1}^N$ is often unknown. More importantly, the impact of local head $\{\boldsymbol{\theta}_i\}_{i=1}^N$ is evaluated by partial gradients of local loss functions $F_i, \forall i$, which is implicitly incorporated in the first two terms in (10), with definitions given in (8).

Remark 2. *The ϵ-approximation solution defined in (9) and* (10) *incorporates the impact of two coupled parameters, while conventional decentralized learning frameworks such as Lian et al. (2017); Assran et al. (2019); Xiong et al. (2023b) only considered a single shared model. This makes existing convergence analysis not directly applicable to ours and necessitates different proof techniques. One fundamental reason is that, instead of learning a single shared model, there are multiple local heads strongly coupled with the global representation that need to be handled in our setting. Compared to the PS framework, only the gap between the learned global representation $\boldsymbol{\phi}$ and the global optimum $\boldsymbol{\phi}^*$ under a linear representation model is considered in Collins et al. (2021). Finally, another line of work on decentralized bilevel optimization (Liu et al. 2022; Qiu et al. 2022) involves two coupled parameters under the assumption that inner parameters are strongly convex in outer parameters, and hence differ from our model and definition in* (10).

Assumptions

Assumption 1 (Doubly Stochastic Consensus Matrix)**.** *The consensus matrix $\mathbf{P} = (P_{i,j})$ is doubly stochastic, i.e., $\sum_{j=1}^N P_{i,j} = \sum_{i=1}^N P_{i,j} = 1, \forall i \in [N], j \in [N]$.*

Assumption 2 (L-Lipschitz Continuous Gradient)**.** *There exists a constant $L > 0$, such that $\|\nabla_{\boldsymbol{\phi}} F_i(\boldsymbol{\phi}, \boldsymbol{\theta}) - \nabla_{\boldsymbol{\phi}} F_i(\boldsymbol{\phi}', \boldsymbol{\theta}')\| \leq L(\|\boldsymbol{\phi} - \boldsymbol{\phi}'\| + \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|)$ and $\|\nabla_{\boldsymbol{\theta}} F_i(\boldsymbol{\phi}, \boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} F_i(\boldsymbol{\phi}', \boldsymbol{\theta}')\| \leq L(\|\boldsymbol{\phi} - \boldsymbol{\phi}'\| + \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|)$, $\forall i \in [N], \forall \boldsymbol{\phi}, \boldsymbol{\phi}' \in \Phi, \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$.*

Assumption 3 (Unbiased Local Gradient Estimator)**.** *The local gradient estimators are unbiased, i.e., $\forall \boldsymbol{\phi}_i, \boldsymbol{\phi}_i' \in \Phi$, $\forall \boldsymbol{\theta}_i, \boldsymbol{\theta}_i' \in \Theta$, $\forall i \in [N]$, $\mathbb{E}[g_{\boldsymbol{\phi}}(\boldsymbol{\phi}_i, \boldsymbol{\theta}_i)] = \nabla_{\boldsymbol{\phi}} F_i(\boldsymbol{\phi}_i, \boldsymbol{\theta}_i)$, $\mathbb{E}[g_{\boldsymbol{\theta}}(\boldsymbol{\phi}_i, \boldsymbol{\theta}_i)] = \nabla_{\boldsymbol{\theta}} F_i(\boldsymbol{\phi}_i, \boldsymbol{\theta}_i)$, with the expectation being taken over the local data samples.*

Assumption 4 (Bounded Variance)**.** *There exists a constant $\sigma > 0$ such that the variance of each local gradient estimator is bounded, i.e., $\forall \boldsymbol{\phi}_i, \boldsymbol{\phi}_i' \in \Phi$, $\forall \boldsymbol{\theta}_i, \boldsymbol{\theta}_i' \in \Theta$, $\forall i \in [N]$, $\mathbb{E}[\|g_{\boldsymbol{\phi}}(\boldsymbol{\phi}_i, \boldsymbol{\theta}_i) - \nabla_{\boldsymbol{\phi}} F_i(\boldsymbol{\phi}_i, \boldsymbol{\theta}_i)\|^2] \leq \sigma^2$, $\mathbb{E}[\|g_{\boldsymbol{\theta}}(\boldsymbol{\phi}_i, \boldsymbol{\theta}_i) - \nabla_{\boldsymbol{\theta}} F_i(\boldsymbol{\phi}_i, \boldsymbol{\theta}_i)\|^2] \leq \sigma^2$.*

Assumption 5 (Bounded Global Variability)**.** *There exists a constant* $\varsigma > 0$ *such that the global variability of the local partial gradients on* $\boldsymbol{\phi}$ *of the loss function* $\forall \boldsymbol{\theta}_i \in \Theta$ *is bounded, i.e.,* $\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|\nabla_{\boldsymbol{\phi}} F_i(\boldsymbol{\phi},\boldsymbol{\theta}_i) - \nabla_{\boldsymbol{\phi}} f(\boldsymbol{\phi},\{\boldsymbol{\theta}_i\}_{i=1}^N)\|^2] \leq \varsigma^2$.

Assumptions 1-5 are standard (Kairouz et al. 2019; Tang et al. 2020; Yang, Fang, and Liu 2021), except the difference caused by two coupled parameters in our representation learning model. We use a universal bound ς to quantify the global variability of local partial gradients on global representation $\boldsymbol{\phi}$ in Assumption 5 due to the non-i.i.d. data among workers, which is similar to the heterogeneity assumption in conventional decentralized frameworks with a single global parameter, where $\|\nabla_{\mathbf{w}} F_i(\mathbf{w}) - \nabla_{\mathbf{w}} f(\mathbf{w})\|^2 \leq \varsigma^2, \forall i \in [N]$ with $\varsigma = 0$ meaning i.i.d. data across workers (Lian et al. 2017). Finally, it is worth noting that we do *not* require a bounded gradient assumption, which is often used in distributed optimization analysis (Nedic and Ozdaglar 2009).

Convergence Analysis for DePRL

Theorem 1. *Under Assumptions 1-5, with learning rates* $\alpha \leq \frac{1+36\tau^2}{\tau L}$ *and* $\beta \leq \min\left(1/L, N/2, \frac{1-q}{3\sqrt{2}CLN}\right)$, *where* $C := \frac{2(1+p^{-N})}{1-p^N}$, $q := (1-p^N)^{1/N}$ *and* $p = \arg\min P_{i,j}, \forall i,j, P_{i,j} > 0$. *Denote the optimal parameters of global representation and local heads as* $\boldsymbol{\phi}^*$ *and* $\{\boldsymbol{\theta}_i^*\}_{i=1}^N$, *respectively. The sequence of parameters* $\{\{\boldsymbol{\phi}_i(k)\}_{i=1}^N, \{\boldsymbol{\theta}_i(k)\}_{i=1}^N, \forall k\}$ *generated by* DePRL *satisfy*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[M(k)] \leq \frac{4f(\bar{\boldsymbol{\phi}}(0),\{\boldsymbol{\theta}_i(0)\}_{i=1}^N) - 4f(\boldsymbol{\phi}^*,\{\boldsymbol{\theta}_i^*\}_{i=1}^N)}{K\beta} + \frac{2\beta L}{N}\sigma^2 + \frac{12\alpha^3 L^2\tau}{\beta}(\tau-1)(6\tau+1)\sigma^2 + \frac{2\alpha^2\tau L}{\beta}\sigma^2 + \frac{2\beta}{3N}\left(1+\frac{1}{L^2}\right)\sigma^2 + \frac{2\beta}{N}\left(1+\frac{1}{L^2}\right)\varsigma^2. \tag{11}$$

There are two terms on right hand side of (11): (i) a vanishing term $\frac{4f(\bar{\boldsymbol{\phi}}(0),\{\boldsymbol{\theta}_i(0)\}_{i=1}^N) - 4f(\boldsymbol{\phi}^*,\{\boldsymbol{\theta}_i^*\}_{i=1}^N)}{K\beta}$ that goes to zero as K increases; and (ii) a constant noise term $\frac{2\beta L}{N}\sigma^2 + \frac{12\alpha^3 L^2\tau}{\beta}(\tau-1)(6\tau+1)\sigma^2 + \frac{2\alpha^2\tau L}{\beta}\sigma^2 + \frac{2\beta}{3N}\left(1+\frac{1}{L^2}\right)\sigma^2 + \frac{2\beta}{N}\left(1+\frac{1}{L^2}\right)\varsigma^2$ that depends on the problem instance and is independent of K. The decay rate of the vanishing term matches that of conventional decentralized learning frameworks with a single shared model (Lian et al. 2017, 2018; Assran et al. 2019). The constant noise term mainly comes from the variance of stochastic partial gradients on consensus global representation $\bar{\boldsymbol{\phi}}(k)$ and local head $\{\boldsymbol{\theta}_i(k)\}_{i=1}^N$, as well as the global variability due to the model heterogeneity when bounding the consensus error of global representation, i.e., $\|\bar{\boldsymbol{\phi}}(k) - \boldsymbol{\phi}_i(k)\|^2, \forall i \in [N]$ at each round k. In particular, the noise term $\frac{12\alpha^3 L^2\tau}{\beta}(\tau-1)(6\tau+1)\sigma^2 + \frac{2\alpha^2\tau L}{\beta}\sigma^2$ is caused by τ-step local head update, where the first part measures the accumulated error between each intermediate update $\boldsymbol{\theta}_i(k,s), \forall s \in \{0,1,\ldots,\tau-1\}$ and $\boldsymbol{\theta}_i(k)$, i.e., $\mathbb{E}\|\boldsymbol{\theta}_i(k,s) - \boldsymbol{\theta}_i(k)\|^2$, and goes to zero when $\tau = 1$ due to the fact that $\boldsymbol{\theta}_i(k,0) = \boldsymbol{\theta}_i(k)$. To lower its impact, an inverse relationship between the local head learning rate α and the updated steps τ in each round is desired, i.e., $\alpha = \mathcal{O}(\frac{1}{\tau})$, such that the error can be offset by a small α. This is consistent with observations in the PS setting with non-IID datasets across workers (Yang, Fang, and Liu 2021).

Corollary 1. *Let* $\alpha = \frac{1}{\tau\sqrt{K}}$ *and* $\beta = \sqrt{N/K}$. *The convergence rate of* DePRL *is* $\mathcal{O}\left(\frac{1}{\sqrt{NK}} + \frac{1}{K\sqrt{N}} + \frac{1}{\tau\sqrt{NK}}\right)$, *when the total number of communication rounds* K *satisfies* $K \geq \max\left(\frac{18C^2L^2N^3}{(1-q)^2}, \frac{(2L^2+2)^2}{NL^4}, NL^2\right)$.

Since $\frac{1}{K\sqrt{N}}$ and $\frac{1}{\tau\sqrt{NK}}$ are dominated by $\frac{1}{\sqrt{NK}}$, DePRL with two coupled parameters achieves a **linear speedup** for convergence, i.e., we can proportionally decrease K as N increases while keeping the same convergence rate. This is the first linear speedup result for personalized decentralized learning with shared representations, and is highly desirable since it implies that one can efficiently leverage the massive parallelism in large-scale decentralized systems. An interesting point is that our result also indicates that the number of local updates τ does not hurt the convergence with a proper learning rate choice for α as observed in PS setting (Yang, Fang, and Liu 2021). Need to mention that local SGD steps usually slow down the convergence around $\mathcal{O}(\frac{\tau}{K})$ even for strongly convex objectives as shown in Li et al. (2020).

Intuitions and Proof Sketch

We now highlight the key ideas and challenges behind the convergence proof of DePRL with two coupled parameters. Given the definition of ϵ-approximation solution defined in (9) and (10), we characterize the descending property of the global loss function as

$$\begin{aligned}
&\mathbb{E}[f(\bar{\boldsymbol{\phi}}(k+1),\{\boldsymbol{\theta}_i(k+1)\}_{i=1}^N)] - \mathbb{E}[f(\bar{\boldsymbol{\phi}}(k),\{\boldsymbol{\theta}_i(k)\}_{i=1}^N)] \\
&\leq \underbrace{\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\langle \nabla_{\boldsymbol{\phi}} F_i(\bar{\boldsymbol{\phi}}(k),\boldsymbol{\theta}_i(k+1)), \bar{\boldsymbol{\phi}}(k+1) - \bar{\boldsymbol{\phi}}(k)\right\rangle}_{C_1} \\
&\quad + \underbrace{\frac{1}{N}\sum_{i=1}^{N}\frac{L}{2}\mathbb{E}[\|\bar{\boldsymbol{\phi}}(k+1) - \bar{\boldsymbol{\phi}}(k)\|^2]}_{C_2} \\
&\quad + \underbrace{\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\langle \nabla_{\boldsymbol{\theta}} F_i(\bar{\boldsymbol{\phi}}(k),\boldsymbol{\theta}_i(k)),\ \boldsymbol{\theta}_i(k+1) - \boldsymbol{\theta}_i(k)\right\rangle}_{C_3} \\
&\quad + \underbrace{\frac{1}{N}\sum_{i=1}^{N}\frac{L}{2}\mathbb{E}[\|\boldsymbol{\theta}_i(k+1) - \boldsymbol{\theta}_i(k)\|^2]}_{C_4},
\end{aligned} \tag{12}$$

by following the Lipschitz assumption. Bounding C_1, C_2, C_3 and C_4 leads to all key components in $M(k)$ defined in (10), including the partial gradients on $\bar{\boldsymbol{\phi}}(k)$ and $\{\boldsymbol{\theta}_i(k)\}_{i=1}^N$ of the global loss function, the error of gradient estimation, and the average consensus error of the global representation.

Dataset (Model)	π	Ring			Random			FC		
		D-PSGD	DisPFL	DePRL	D-PSGD	DisPFL	DePRL	D-PSGD	DisPFL	DePRL
CIFAR-100 (ResNet-18)	0.1	25.18±0.4	46.09±0.2	**60.72**±0.2	30.27±0.5	47.77±0.3	**61.51**±0.5	33.04±0.7	47.96±0.3	**62.40**±0.7
	0.3	26.51±0.4	37.92±0.3	**49.82**±0.4	31.71±0.3	39.91±0.5	**50.49**±0.7	34.93±0.7	40.37±0.5	**51.51**±0.5
	0.5	26.90±0.3	35.33±0.4	**45.89**±0.4	32.05±0.4	37.39±0.3	**46.68**±0.4	35.22±0.6	37.86±0.3	**47.63**±0.3
CIFAR-10 (VGG-11)	0.1	53.91±0.2	86.38±0.3	**89.57**±0.2	57.69±0.2	89.01±0.2	**91.03**±0.1	58.90±0.1	89.19±0.3	**91.33**±0.2
	0.3	59.17±0.2	73.48±0.3	**76.41**±0.3	64.12±0.4	77.36±0.4	**79.60**±0.2	65.82±0.3	78.52±0.5	**79.84**±0.4
	0.5	60.45±0.4	68.83±0.2	**72.51**±0.2	65.48±0.2	73.30±0.4	**74.80**±0.2	67.30±0.3	74.50±0.3	**75.04**±0.2
Fashion MNIST (AlexNet)	0.1	77.45±0.2	95.74±0.2	**96.66**±0.2	84.74±0.2	96.59±0.3	**97.16**±0.2	87.24±0.3	96.76±0.3	**97.36**±0.2
	0.3	81.95±0.5	91.52±0.3	**92.81**±0.2	87.76±0.3	93.47±0.3	**94.81**±0.2	89.80±0.5	93.50±0.2	**95.03**±0.2
	0.5	84.63±0.2	89.49±0.2	**91.36**±0.3	88.93±0.5	91.99±0.3	**93.55**±0.2	90.38±0.4	92.13±0.2	**93.87**±0.3
HARBox (DNN)	0.1	54.90±0.7	90.96±0.1	**92.07**±0.1	57.59±0.6	91.36±0.3	**92.49**±0.1	58.23±0.3	91.47±0.1	**93.46**±0.1
	0.3	55.41±0.7	80.02±0.2	**80.85**±0.1	57.97±0.7	80.35±0.2	**81.30**±0.2	58.93±0.7	82.14±0.2	**83.55**±0.2
	0.5	56.66±0.7	74.47±0.1	**75.84**±0.1	58.59±0.7	74.72±0.3	**76.22**±0.2	59.17±0.7	77.61±0.3	**78.74**±0.2

Table 1: Average test accuracy with different communication graphs and data heterogeneities.

As aforementioned, instead of learning a single shared model as in conventional decentralized learning frameworks (Lian et al. 2017, 2018), DePRL needs to handle multiple local heads that are strongly coupled with the global representation, which necessitates different proof techniques. Below, we highlight several key differences: 1) ***Coupled model parameters.*** The updates of local heads $\{\boldsymbol{\theta}_i\}_{i=1}^N$ and global representations $\{\boldsymbol{\phi}_i\}_{i=1}^N$ are strongly coupled, which makes bounding C_1 and C_3 challenging. In particular, the update of consensus global representation $\bar{\boldsymbol{\phi}}(k)$ depends on local heads $\{\boldsymbol{\theta}_i(k+1)\}_{i=1}^N$ in C_1, and the update of local heads $\{\boldsymbol{\theta}_i(k)\}_{i=1}^N$ depends on the consensus global representation $\bar{\boldsymbol{\phi}}(k)$ in C_3. Since the loss function is evaluated on consensus global representation $\bar{\boldsymbol{\phi}}(k)$, we show that bounding C_1 and C_3 can be reduced to bound the consensus error of global representation $\|\boldsymbol{\phi}_i(k) - \bar{\boldsymbol{\phi}}(k)\|^2$. Specifically, the bound on consensus error allows us to control the terms involving the local partial gradients and local updates in the drift of the global loss function as shown in (12), which also serves as a bridge to track the update of $\bar{\boldsymbol{\phi}}(k+1) - \bar{\boldsymbol{\phi}}(k+1)$ in C_2 and $\boldsymbol{\theta}_i(k+1) - \boldsymbol{\theta}_i(k)$ in C_4. 2) ***Consensus error.*** Based on (5), (6) and (7), the average consensus error $\mathbb{E}\|\boldsymbol{\phi}_i(k) - \bar{\boldsymbol{\phi}}(k)\|^2$ depends on both the consensus matrix $\mathbf{P}$, and the local partial gradient on $\boldsymbol{\phi}$, i.e., $g_{\boldsymbol{\phi}}(\boldsymbol{\phi}_i(k), \boldsymbol{\theta}_i(k+1))$, which is correlated with local heads $\{\boldsymbol{\theta}_i(k)\}_{i=1}^N$. We address these impacts by leveraging Assumption 5. 3) ***Two learning rates.*** We leverage a weight term in (10) to capture the different learning rates for $\{\boldsymbol{\theta}_i(k)\}_{i=1}^N$ and $\{\boldsymbol{\phi}_i(k)\}_{i=1}^N$. This weight benefits for characterizing the desired learning rate for convergence.

Experiments

We experimentally evaluate the performance of DePRL . Further details about experiments, hyperparameters, and additional results are provided in Xiong et al. (2023a).

Datasets and Models. We use (i) three image classification datasets: CIFAR-100, CIFAR-10 (Krizhevsky, Hinton et al. 2009) and Fashion-MNIST (Xiao, Rasul, and Vollgraf 2017); and (ii) a human activity recognition dataset: HARBox (Ouyang et al. 2021). We simulate non-IID scenario by considering a heterogeneous data partition for which the number of data points and class proportions are unbalanced Wang et al. (2020a,b). In particular, we simulate a heterogeneous partition into N workers by sampling $\boldsymbol{p}_i \sim \text{Dir}_N(\pi)$, where π is the parameter of Dirichlet distribution. We use ResNet-18 (He et al. 2016) for CIFAR-100, VGG-11 (Simonyan and Zisserman 2015) for CIFAR-10, AlexNet (Krizhevsky, Sutskever, and Hinton 2012) for Fashion-MNIST, and a fully connected DNN (Li et al. 2021a, 2022). As in Collins et al. (2021), we treat the head as the weights and biases as the final fully-connected layer in each of the models.

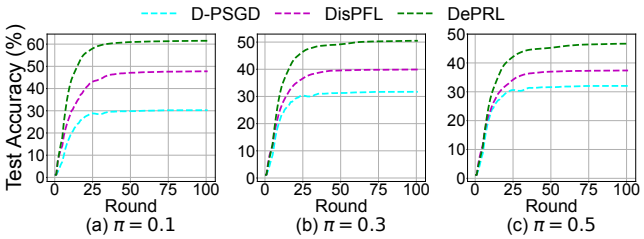

Figure 2: Learning curves of different baselines using ResNet-18 on non-IID partitioned CIFAR-100 with different heterogeneities when the communication graph is "Random".

Baselines. For decentralized setting, we take the commonly used D-PSGD (Lian et al. 2017) and DisPFL (Dai et al. 2022), a personalized method with a single shared model. PS based baselines include FedAvg (McMahan et al. 2017), FedRep (Collins et al. 2021), Ditto (Li et al. 2021b) and FedRoD (Chen and Chao 2022). We implement all algorithms in PyTorch (Paszke et al. 2017) on Python 3 with three NVIDIA RTX A6000 GPUs.

Communication Graph. Based on our model and theoretical analysis, we randomly generate a connected communication graph ("Random" for short) for decentralized settings. We also experiment on two representative communication graph including "Ring" and "fully connected (FC)". Further, since communications occur between the central server and workers in PS based setting, for a fair comparison (from the perspective of total communications), we only compare decentralized baselines with PS based baselines under "Ring". Due to space constraints, we relegate the comparisons with PS based methods to Xiong et al. (2023a).

Dataset (Model)	π	Ring			Random			FC		
		D-PSGD	DisPFL	DePRL	D-PSGD	DisPFL	DePRL	D-PSGD	DisPFL	DePRL
CIFAR-100 (ResNet-18)	0.1	38.64±0.1	34.67±0.2	**53.58**±0.3	49.81±0.2	42.93±0.2	**53.78**±0.2	51.32±0.2	47.50±0.1	**54.16**±0.2
	0.3	27.75±0.2	24.89±0.1	**44.62**±0.2	42.27±0.1	34.05±0.1	**44.79**±0.1	43.42±0.2	39.20±0.1	**45.48**±0.2
	0.5	26.15±0.2	22.93±0.1	**42.27**±0.2	39.25±0.1	31.11±0.2	**42.65**±0.2	41.20±0.1	37.26±0.1	**42.68**±0.1
CIFAR-10 (VGG-11)	0.1	73.81±0.2	70.85±0.3	**75.93**±0.2	74.78±0.3	73.62±0.5	**80.22**±0.2	75.39±0.3	73.71±0.2	**80.48**±0.2
	0.3	57.34±0.3	51.89±0.4	**59.01**±0.4	58.86±0.2	57.08±0.3	**65.60**±0.3	59.68±0.2	57.18±0.4	**65.67**±0.3
	0.5	50.92±0.2	47.36±0.2	**52.72**±0.2	52.51±0.3	49.75±0.3	**63.04**±0.2	53.67±0.2	49.81±0.2	**63.29**±0.3
Fashion MNIST (AlexNet)	0.1	84.76±0.3	83.72±0.3	**87.34**±0.2	85.37±0.4	85.44±0.3	**88.48**±0.2	86.48±0.3	85.65±0.3	**88.86**±0.2
	0.3	74.84±0.3	73.07±0.2	**78.29**±0.3	78.54±0.3	76.92±0.2	**80.74**±0.2	79.16±0.2	77.69±0.2	**80.84**±0.2
	0.5	67.54±0.3	65.64±0.4	**71.14**±0.3	71.97±0.2	70.31±0.2	**77.37**±0.3	72.57±0.4	70.61±0.2	**77.52**±0.2
HARBox (DNN)	0.1	51.07±0.7	50.58±0.6	**55.97**±0.3	51.23±0.7	51.12±0.7	**56.39**±0.5	51.84±0.7	51.21±0.6	**57.70**±0.3
	0.3	49.50±0.3	48.97±0.3	**55.86**±0.3	49.81±0.3	49.49±0.4	**56.11**±0.3	51.53±0.5	49.55±0.5	**56.39**±0.3
	0.5	48.24±0.3	48.32±0.3	**52.42**±0.3	48.28±0.3	48.47±0.2	**52.46**±0.3	48.82±0.2	48.49±0.2	**52.59**±0.3

Table 2: Generalization performance in terms of test accuracy.

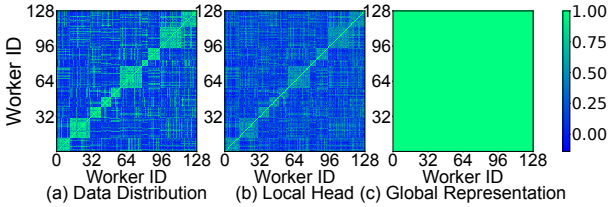

Figure 3: Similarities between workers on (a) data distributions; (b) local heads; and (c) global representation.

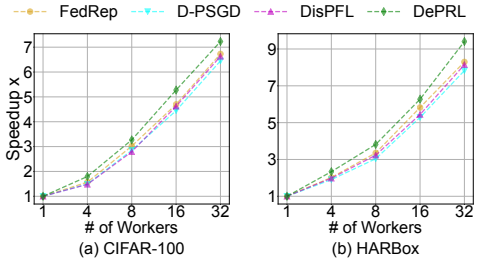

Figure 4: Speedup with different number of workers.

Configurations. All results are averaged over four random seeds. The final accuracy is calculated through the average of each worker's local test accuracy. The total worker number is 128, and the epoch number for local head update is 2. An ablation study is conducted in Xiong et al. (2023a).

Testing Accuracy. We show the final test accuracy for all considered algorithms under various settings in Table 1, and report the learning curve in Figure 2. First, the state-of-the-art D-PSGD performs worse in non-IID settings due to the fact that it targets on learning a single model without encouraging personalization. Second, though DisPFL is incorporated with personalization, and significantly improves the performance of D-PSGD, DePRL always outperforms DisPFL. In particular, DePRL achieves a remarkable performance improvement on non-IID partitioned CIFAR-100. Compared to CIFAR-10, the data heterogeneity across workers are further increased due to the larger number of classes, and hence calls for personalization of local models. This observation makes our representation learning augmented personalized model in DePRL even pronounced compared to learning a single full-dimensional model in these baseline methods. Finally, the superior performance of DePRL over D-PSGD and DisPFL is consistent and robust over all communication graphs.

Learned Local Head and Global Representation. To further advocate the benefits of DePRL for producing personalized models via leveraging representation learning theory, we report the distance between learned local heads, global representation and task similarities. We measure the similarities by cos-similarity between data distributions, learned local heads, and learned global representation across workers. As shown in Figure 3 on non-IID partitioned CIFAR-100, DePRL is able to accommodate the heterogeneities among workers not only with the learned local heads in alignment with local data distribution, but also with the same global representation. This further validates our theoretical analysis.

Speedup. We validate our main theoretical result that DePRL achieves a linear speedup of convergence. This often means that the convergence time (measures how quickly the gradient norm converges to zero) should be linear in the number of workers. Specifically, we measure the convergence time of different algorithms, and compute their speedup with respect to that of a centralized setting as in D-PSGD (Lian et al. 2017). From Figure 4, we observe that the speedup (convergence time) is almost linearly increasing (decreasing) as the number of workers increases.

Generalization to New Workers. We evaluate the effectiveness of global representation learned by DePRL when generalizes it to new workers. Specifically, when training all considered models using different datasets with $\alpha = 0.3$, all initial 128 workers collaboratively learn a corresponding global representation $\boldsymbol{\phi}$. Then we encounter 64 new workers and partition the datasets across these new workers, which lead to significantly different datasets across workers compared to all initial workers. Each new worker i leverages the learned global representation by initial workers, and perform multiple local steps to learn its local head $\boldsymbol{\theta}_i$. We then evaluate the test accuracy as above across these 64 new workers. As shown in Table 2, DePRL significantly outperforms all baselines in the generalization performance across all communication graphs and varying levels of heterogeneity.

Acknowledgements

This work was supported in part by the National Science Foundation (NSF) grants 2148309 and 2315614, and was supported in part by funds from OUSD R&E, NIST, and industry partners as specified in the Resilient & Intelligent NextG Systems (RINGS) program. This work was also supported in part by the U.S. Army Research Office (ARO) grant W911NF-23-1-0072, and the U.S. Department of Energy (DOE) grant DE-EE0009341. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Assran, M.; Loizou, N.; Ballas, N.; and Rabbat, M. 2019. Stochastic gradient push for distributed deep learning. In *International Conference on Machine Learning*, 344–353. PMLR.

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

Borkar, V. S. 2009. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer.

Chen, H.-Y.; and Chao, W.-L. 2022. On Bridging Generic and Personalized Federated Learning for Image Classification. In *ICLR*.

Collins, L.; Hassani, H.; Mokhtari, A.; and Shakkottai, S. 2021. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, 2089–2099. PMLR.

Dai, R.; Shen, L.; He, F.; Tian, X.; and Tao, D. 2022. DisPFL: Towards Communication-Efficient Personalized Federated Learning via Decentralized Sparse Training. In *International Conference on Machine Learning*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proc. of IEEE CVPR*.

Imteaj, A.; Mamun Ahmed, K.; Thakker, U.; Wang, S.; Li, J.; and Amini, M. H. 2022. Federated learning for resource-constrained IoT devices: panoramas and state of the art. *Federated and Transfer Learning*, 7–27.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2019. Advances and Open Problems in Federated Learning. *arXiv preprint arXiv:1912.04977*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning Multiple Layers of Features from Tiny Images.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet Classification with Deep Convolutional Neural Networks. *Proc. of NIPS*.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature*, 521(7553): 436–444.

Li, A.; Sun, J.; Li, P.; Pu, Y.; Li, H.; and Chen, Y. 2021a. Hermes: an efficient federated learning framework for heterogeneous mobile clients. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, 420–437.

Li, C.; Zeng, X.; Zhang, M.; and Cao, Z. 2022. PyramidFL: A fine-grained client selection framework for efficient federated learning. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, 158–171.

Li, T.; Hu, S.; Beirami, A.; and Smith, V. 2021b. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, 6357–6368. PMLR.

Li, X.; Huang, K.; Yang, W.; Wang, S.; and Zhang, Z. 2020. On the Convergence of FedAvg on Non-IID Data. In *International Conference on Learning Representations*.

Lian, X.; Zhang, C.; Zhang, H.; Hsieh, C.-J.; Zhang, W.; and Liu, J. 2017. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30.

Lian, X.; Zhang, W.; Zhang, C.; and Liu, J. 2018. Asynchronous Decentralized Parallel Stochastic Gradient Descent. In *Proc. of ICML*.

Liu, Z.; Zhang, X.; Khanduri, P.; Lu, S.; and Liu, J. 2022. INTERACT: achieving low sample and communication complexities in decentralized bilevel learning over networks. In *Proceedings of the Twenty-Third International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, 61–70.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-Efficient Learning of Deep Networks From Decentralized Data. In *Proc. of AISTATS*.

Nedić, A.; Olshevsky, A.; and Rabbat, M. G. 2018. Network Topology and Communication-Computation Tradeoffs in Decentralized Optimization. *Proceedings of the IEEE*, 106(5): 953–976.

Nedic, A.; and Ozdaglar, A. 2009. Distributed Subgradient Methods for Multi-Agent Optimization. *IEEE Transactions on Automatic Control*, 54(1): 48–61.

Ouyang, X.; Xie, Z.; Zhou, J.; Huang, J.; and Xing, G. 2021. Clusterfl: a similarity-aware federated learning system for human activity recognition. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, 54–66.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Qiu, P.; Li, Y.; Liu, Z.; Khanduri, P.; Liu, J.; Shroff, N. B.; Bentley, E. S.; and Turck, K. 2022. DIAMOND: Taming Sample and Communication Complexities in Decentralized Bilevel Optimization. *arXiv preprint arXiv:2212.02376*.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-scale Image Recognition. In *Proc. of ICLR*.

Tang, Z.; Shi, S.; Chu, X.; Wang, W.; and Li, B. 2020. Communication-Efficient Distributed Deep Learning: A Comprehensive Survey. *arXiv preprint arXiv:2003.06307*.

Vanhaesebrouck, P.; Bellet, A.; and Tommasi, M. 2017. Decentralized collaborative learning of personalized models over networks. In *Artificial Intelligence and Statistics*, 509–517. PMLR.

Wang, H.; Yurochkin, M.; Sun, Y.; Papailiopoulos, D.; and Khazaeni, Y. 2020a. Federated Learning with Matched Averaging. In *Proc. of ICLR*.

Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; and Poor, H. V. 2020b. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization. *Proc. of NeurIPS*.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*.

Xiong, G.; Yan, G.; Wang, S.; and Li, J. 2023a. DePRL: Achieving Linear Convergence Speedup in Personalized Decentralized Learning with Shared Representations. *arXiv preprint arXiv:2312.10815*.

Xiong, G.; Yan, G.; Wang, S.; and Li, J. 2023b. Straggler-Resilient Decentralized Learning via Adaptive Asynchronous Updates. *arXiv preprint arXiv:2306.06559*.

Yang, H.; Fang, M.; and Liu, J. 2021. Achieving Linear Speedup with Partial Worker Participation in Non-IID Federated Learning. In *Proc. of ICLR*.