Exploiting Feature Heterogeneity for Improved Generalization in Federated Multi-task Learning

Renpu Liu*, Jing Yang*, and Cong Shen†
*The Pennsylvania State University, University Park, PA 16802, USA
†University of Virginia, Charlottesville, VA 22904, USA

***Abstract*—In this work, we investigate a general federated multi-task learning (FMTL) problem where each task may be performed at multiple clients, and each client may perform multiple tasks. Although the tasks share some common representation (i.e., feature-map) that can help to learn, the distribution of the features in the feature space may vary across different tasks at different clients, which poses a significant challenge to FMTL. While non-independent and identically distributed (non-IID) local datasets at different clients are often considered detrimental to model convergence in federated learning (FL), such statistical heterogeneity in feature space may be beneficial to the generalization performance. In this work, we establish the impact of statistical feature heterogeneity on generalization, through the lens of a multi-task linear regression model. In order to leverage the feature distribution heterogeneity, we propose a novel augmented dataset based approach, and prove that under certain conditions, FMTL on heterogeneous datasets can outperform the homogeneous counterpart in terms of the generalization performance. The theoretical analysis further leads to a simple client weighting method based on optimizing the excess risk upper bound. Experimental results demonstrate that the generalization performance can be improved on a real-world dataset with the proposed method.**

I. Introduction

Federated learning (FL) [1] is a novel distributed machine learning (ML) paradigm where a massive number of clients are orchestrated to train a global ML model collaboratively while keeping all the training data on local devices [2]. Among the most important characteristics of FL is the non-independent and identically distributed (non-IID) local datasets. While the majority of the existing literature focuses on the convergence [3]–[5] and communication [6] implications of such distributional heterogeneity, only a few works investigate its impact on the generalization performance of FL [7], [8]. In [7], the authors propose a new agnostic federated learning framework where the objective is to obtain a global model that minimizes the maximum loss on any target distribution formed by a mixture of the client distributions. The generalization performance is characterized by a data-dependent Rademacher complexity. In [9], it assumes that clients are drawn from a meta-distribution, with their data drawn from local data distributions. Thus the generalization performance depends on both unseen client data and unseen client distributions. However, the explicit relationship between data distribution and generalization is not explicitly characterized.

In this work, we make an initial effort to investigate the impact of data distributional heterogeneity on the generalization of FL. In stark contrast to the conventional belief that data heterogeneity would compromise the performance of FL, we are able to show that the heterogeneity in feature space can be utilized to improve the generalization instead. Toward this end, we focus on a novel federated multi-task learning (FMTL) model. Specifically, in our setting, we assume that each client may perform multiple regression tasks, and each task may be performed by multiple clients. While data collected under different tasks share a common representation (i.e., feature-map) [10]–[14], the corresponding linear regression model varies across tasks. This naturally induces *inter-task* data distribution heterogeneity. Meanwhile, the total number of samples collected under a single task varies among clients, and different clients could collect data for a single task from different distributions, leading to the *intra-task* heterogeneity. Our objective is to theoretically characterize how these two different types of heterogeneity influence the excess risk, and how to adjust the model aggregation weights for each client in order to improve the generalization performance.

We note that the shared representation among tasks is a common assumption in multi-task learning [10]–[14] and personalized FL [15]–[23], where task/domain-specific predictors are used on top of the common representation function to model the input-output relations under different tasks/domains. While the focus of those works is to learn the shared representation and the personalized predictor heads, in this work, we focus on the impact of inter-task and intra-task feature distribution heterogeneity on the FL generalization performance. Therefore, we begin with a *known* common representation, and explicitly analyze how the feature heterogeneity influences the corresponding excess risk upper bound.

Our main contributions are three-fold:

• First, we introduce a general FMTL model where multiple clients collaboratively perform multiple learning tasks through the coordination of a parameter server. Although the data samples for different tasks share a common representation, the distributions of features in the feature space vary across tasks and clients, leading to inter-task and intra-task heterogeneity,

The work of RL and JY was supported in part by the US National Science Foundation (NSF) under awards CNS-195627, CNS-2114542, ECCS-2133170, and the US Department of Energy (DOE) under award DE-EE0008763. The work of CS was supported in part by the US NSF under awards ECCS-2033671, ECCS-2143559, and Virginia Commonwealth Cyber Initiative Innovation and Commercialization Award VV-1Q23-005.

respectively.

• Second, in order to leverage such feature distribution heterogeneity, we propose a novel *augmented dataset*-based approach to solving the federated multi-task linear regression problem. Such augmented datasets essentially increase the effective sample size for each task, and potentially improve the generalization performance of the trained model. We then explicitly characterize the generalization performance of the trained model and provide distribution-dependent excess risk upper bounds. Our results indicate that under certain assumptions, the excess risk bound is improved compared with a centralized single-task linear regression problem with the same total amount of IID samples, and the acceleration can be up to $\mathcal{O}(1/m)$, where m is the number of tasks performed at each client.

• Third, based on an explicit excess risk upper bound, we propose a distribution-free weight assignment method for model aggregation. We then extend the weight assignment method derived from the linear regression model to more general regression models and validate the effectiveness of this approach through experiments on the standard CIFAR-10 and CIFAR-100 datasets [24]. Our experimental results indicate that the global model trained from our algorithm outperforms other global models trained from classic FL algorithms for about $\sim 10\%$ on CIFAR-10 and $\sim 7\%$ on CIFAR-100.

Notations. Throughout this paper, bold capital letters (e.g., $\mathbf{X}$) denote matrices, and calligraphic capital letters (e.g., $\mathcal{C}$) denote sets. We use $\mathrm{tr}(\mathbf{X})$ to denote the trace of matrix $\mathbf{X}$, $\sigma_{\min}(\mathbf{X})$ and $\sigma_{\max}(\mathbf{X})$ to denote the minimum and maximum singular value of $\mathbf{X}$ separately. We use $\mathrm{diag}(x_1, \cdots, x_d)$ to denote a d-dimension diagonal matrix with diagonal entries $x_1, \cdots, x_d$, and $|\mathcal{C}|$ denotes the cardinality of set $\mathcal{C}$. We use $\langle x, y\rangle$ to denote the inner product of x and y, use $\|x\|$ to denote the Euclidean norm of vector x, and use $\|x\|^2_{\mathbf{\Lambda}}$ to denote $x^T\mathbf{\Lambda}x$. For matrix $\mathbf{X}$, $\|\mathbf{X}\|$ denotes $\sigma_{\max}(\mathbf{X})$. For $N \in \mathbb{N}$, $[N]$ denotes the set $\{1, \cdots, N\}$. We use $x \sim P_X$ to denote that x is randomly drawn according to distribution P_X.

II. Problem Formulation

We consider a general FMTL system consisting of N clients, indexed by $i \in [N]$, and one central server. We assume there are M different prediction tasks in the system, and each client i may perform $m_i \leq M$ of them. For simplicity, we assume $m_i = m$ for any client i in this work. We use $\mathcal{C}_i \subseteq [M]$ to denote the subset of tasks performed at client i. Besides, we also assume that the same task c may be performed at different clients. We use $\mathcal{I}^c = \{i : c \in \mathcal{C}_i\}$ to denote the set of clients that perform task c.

Let $\mathcal{D}_i^c$ be the collection of training samples for task $c \in \mathcal{C}_i$ at client i, and $n_i^c := |\mathcal{D}_i^c|$. Then, in total, there are $n_i = \sum_{c\in\mathcal{C}_i} n_i^c$ training samples at client i, and there are $n^c = \sum_{i=1}^N n_i^c$ total samples for task c from all clients. Let $n = \sum_{i\in[N]} n_i$. Let (x, y) be a training pair in $\mathcal{D}_i^c$. Then, $(x, y) \sim P_{i,X}^c P_{Y|X}^c$, where $P_{i,X}^c$ is the local data distribution for client i under task c, and $P_{Y|X}^c$ is the conditional distribution of label

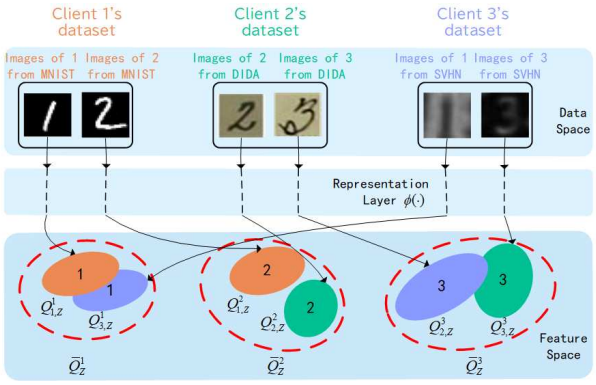

Fig. 1. A toy example of a $N = 3$, $M = 3$ and $m = 2$ model. Each individual client's dataset contains images of 2 kinds of digits, drawn from MNIST [25], DIDA [26], and SVHN [27] separately. The solid ellipses with filled colors refer to each task's feature distribution of each client, and the red dashed ellipses represent the agnostic ground-truth feature distribution of each task.

Y conditioning on data X under task c. We note that $P_{Y|X}^c$ depends on task c only, and is homogeneous across the clients.

We assume there exists a common representation ϕ that maps each raw feature vector x to a feature space. Such a feature map is shared across all tasks. Let $z = \phi(x)$. Then, the distribution of x induces the distribution of z in the feature space. Let the corresponding feature distribution induced by $P_{i,X}^c$ as $Q_{i,Z}^c$.

We note that the heterogeneous data distribution introduces two types of heterogeneity in the feature space. First, as $P_{i,X}^c$ varies across tasks c, the corresponding $Q_{i,Z}^c$ is different among tasks, which we term as *inter-task* heterogeneity. Second, even for the same task c, $P_{i,X}^c$, as well as $Q_{i,X}^c$, may vary across clients i, owing to different data sources of the clients. This leads to so-called *intra-task* distributional heterogeneity in the feature space. We illustrate these two types of heterogeneity in Figure 1.

FMTL Formulation. Once data is embedded in the same feature subspace through ϕ, we can deploy a federated multi-task learning methodology for training from this representation space. We focus on an empirical risk minimization formulation, where the objective function is defined as

$$\hat{L}(\{w_c\}) = \frac{1}{N}\sum_{i\in[N]}\sum_{c\in\mathcal{C}_i}\frac{\beta_i^c}{|\mathcal{D}_i^c|}\sum_{(x,y)\in\mathcal{D}_i^c}\ell\big(f_{w_c}\circ\phi(x), y\big), \quad (1)$$

where β_i^c is the weight assigned to client i for task c, f_{w_c} is a task-specific predictor parameterized by w_c that maps feature vector to the label space, $\circ$ stands for functional composition, and $\ell(\cdot,\cdot)$ is a loss function that measures the error between the true label y and the predicted label $f_{w_c}\circ\phi(x)$.

Excess Risk. In order to measure the generalization performance of the trained model, we define the population risk for any given ϕ as follows:

$$L(\{w_c\}) \coloneqq \sum_{c\in[M]}\alpha^c\mathbb{E}_{\substack{x\sim\bar{P}_X^c\\ y\sim P_{Y|X}^c}}\Big[\ell\big(f_{w_c}\circ\phi(x), y\big)\Big], \quad (2)$$

where α^c is an agnostic weight assigned for task c, and $\bar{P}_X^c$ is the population distribution of the raw features under task c, which may be different from $P_{i,X}^c$, $\forall i$, in general.

Let $\{w_c^*\}$ be the minimizer of the population risk function in (2), and $\{\hat{w}_c\}$ be the minimizer of (1). Then, the excess risk is defined as $L(\{\hat{w}_c\}) - L(\{w_c^*\})$. While characterizing the excess risk for a general learning problem seems intractable, in the following, we focus on multi-task linear regression, which enables us to explicitly bound the corresponding excess risk in a closed form.

III. Multi-task Linear Regression

We consider the following linear regression model. For raw feature $x \in \mathbb{R}^d$ under task c, we assume its true label y is generated according to $y = x^T \mathbf{B} w_c^* + \epsilon$, where $\mathbf{B} \in \mathbb{R}^{d \times k}$ is the feature map and $w_c^* \in \mathbb{R}^k$ is a task-specific linear head. Then, $z = \mathbf{B}^T x$.

Let $\mathbf{X}_i^c$ be a $d \times n_i^c$ matrix whose columns are the raw feature vectors of the data samples included in $\mathcal{D}_i^c$, and Y_i^c be the corresponding vector of labels. We denote $\mathbf{Z}_i^c := \mathbf{B}^T \mathbf{X}_i^c$.

We consider the quadratic loss $\ell(w, z; y) = (w^T z - y)^2$. Then, the empirical and population risks can be written as

$$\hat{L}(\{w_c\}) = \frac{1}{N} \sum_{i \in [N]} \sum_{c \in \mathcal{C}_i} \frac{\beta_i^c}{n_i^c} \left\| (\mathbf{Z}_i^c)^T w_c - Y_i^c \right\|^2, \tag{3}$$

$$L(\{w_c\}) = \sum_{c \in [M]} \alpha^c \mathbb{E}_{\substack{z \sim \bar{Q}_Z^c, \\ \epsilon \sim P_{\mathcal{E}}}} [(w_c^T z - (w_c^*)^T z - \epsilon)^2], \tag{4}$$

respectively, where $\bar{Q}_Z^c$ is the distribution of feature vectors induced by $\bar{P}_X^c$, and $P_{\mathcal{E}}$ is the noise distribution.

Assumptions. We make the following assumptions.

Assumption 1: ϵ is σ_ϵ^2-sub-gaussian, i.e., $\mathbb{E}[\epsilon] = 0$ and $\mathbb{E}\left[\exp(\lambda \epsilon)\right] \leq \exp\left(\frac{\lambda^2 \sigma_\epsilon^2}{2}\right)$ for all $\lambda \in \mathbb{R}$.

Assumption 2 ([28]): Let $z \sim Q_{i,Z}^c$. Then, for any $i \in [N]$ and $c \in \mathcal{C}_i$, z is a sub-gaussian random vector with sub-gaussian norm K, i.e., for all $a \in \mathbb{R}^k$, $\langle z, a \rangle$ is a sub-gaussian random variable and for any $b \in \mathcal{S}^{k-1}$ where $\mathcal{S}^{k-1}$ is the unit sphere defined in $\mathbb{R}^k$, $\langle z, b \rangle$ is a sub-gaussian random variable with sub-gaussian norm K.

In Assumption 2, we follow the definition of sub-gaussian random variables in Lemma 5.5 of [28]. Note that this sub-gaussian definition implies that z is not necessarily a centered random vector. Both Assumptions 1 and 2 are standard in the literature. Next, we present the following feature heterogeneity conditions.

Assumption 3: Let $z \sim Q_{i,Z}^c$. Then, $\mathbf{\Lambda}_i^c := \mathbb{E}[z z^T]$ is invertible for any $i \in [N]$ and $c \in \mathcal{C}_i$. Besides, $\mathbf{Z}_i^c (\mathbf{Z}_i^c)^T$ is invertible almost surely.

Assumption 3 indicates that the energy of z spans all directions of the feature space, while $n_i^c \gg k$ will ensure that $\mathbf{Z}_i^c (\mathbf{Z}_i^c)^T$ is almost surely invertible.

Assumption 4: There exists $\Delta \geq 0$ such that for $z \sim Q_{i,Z}^{c'}$, $c' \neq c$, we have $\mathbb{E}\left[\left(z^T w_c^*\right)^2\right] \leq \Delta$.

When Δ is small, Assumption 4 indicates that the feature vector for task c' is "near-orthogonal" to the predictor head w_c^* associated with task c. Intuitively, for $z \sim Q_{i,Z}^c$, in order to capture its variation and provide an accurate prediction for the corresponding y, the corresponding task-specific predictor w_c^* should be well aligned with it, i.e., $\mathbb{E}\left[\left(z^T w_c^*\right)^2\right]$ should be relatively large. Assumption 4 thus implies a heterogeneous scenario where the distributions of feature vectors from different tasks are well separated in the feature space, where Δ measures the level of feature heterogeneity. We illustrate such feature heterogeneity in Figure 2.

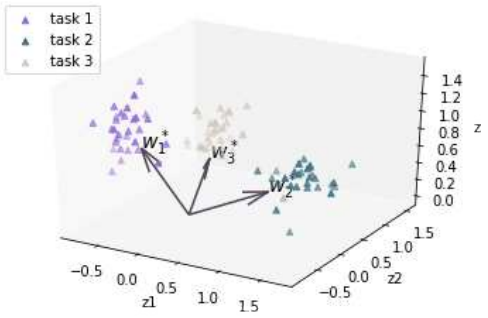

Fig. 2. Feature vectors associated with three tasks. w_1^*, w_2^*, and w_3^* are the corresponding linear heads.

Augmented Datasets. In order to leverage the feature heterogeneity captured by Assumption 4, we propose a novel augmented dataset-based approach. Specifically, for every data sample $(x, y) \in \mathcal{D}_i^c$, we add another sample $(x, 0)$ to $\mathcal{D}_i^{c'}$ for every $c' \neq c$, $c' \in [M]$. We denote the augmented dataset with the added samples as $\tilde{\mathcal{D}}_i^c$. Then, $|\tilde{\mathcal{D}}_i^c| = n_i$ for any $c \in [M]$. Let $\tilde{\mathbf{X}}_i^c \in \mathbb{R}^{d \times n_i}$, $\tilde{Y}_i^c \in \mathbb{R}^{n_i}$, $\tilde{\mathbf{Z}}_i^c \in \mathbb{R}^{k \times n_i}$ be the corresponding raw feature matrix, label vector, and feature matrix defined on the augmented dataset $\tilde{\mathcal{D}}_i^c$ respectively. Then, the new empirical risk function can be expressed as

$$\hat{L}(\{w_c\}) = \frac{1}{nN} \sum_{i \in [N]} \sum_{c \in [M]} \frac{\beta_i^c}{\hat{\beta}_i} \left\| (\tilde{\mathbf{Z}}_i^c)^T w_c - \tilde{Y}_i^c \right\|^2, \tag{5}$$

where $\hat{\beta}_i = n_i / n$. Let $\{\hat{w}_c\}$ be the minimizer of (5). Then, the excess risk can be expressed as

$$L(\{\hat{w}_c\}) - L(\{w_c^*\}) = \sum_{c \in [M]} \alpha^c \|\hat{w}_c - w_c^*\|_{\bar{\mathbf{\Lambda}}^c}^2, \tag{6}$$

where $\bar{\mathbf{\Lambda}}^c := \mathbb{E}_{z \sim \bar{Q}_Z^c}\left[z z^T\right]$.

Note that the empirical risk optimization in (5) can be decomposed into the summation of M independent sub-problems, one associated with each task c. Thus, to evaluate the excess risk in (6), it is equivalent to assessing $\|\hat{w}_c - w_c^*\|_{\bar{\mathbf{\Lambda}}^c}^2$ for each $c \in [M]$, where $\hat{w}_c$ is the minimizer of

$$\hat{L}^c(w_c) = \frac{1}{nN} \sum_{i \in [N]} \frac{\beta_i^c}{\hat{\beta}_i} \left\| (\tilde{\mathbf{Z}}_i^c)^T w_c - \tilde{Y}_i^c \right\|^2. \tag{7}$$

IV. Excess risk upper bound

In this section, we analyze the FMTL generalization performance in terms of the excess risk upper bound for the linear regression model under different heterogeneity assumptions.

A. Both inter- and intra-task heterogeneity

We first consider the general case that $Q_{i,Z}^c$ may vary for different client i and different c, and the sizes of local datasets are not necessarily equal, i.e., $n_i \neq n_j$ in general. To further capture the feature heterogeneity in FMTL, we define $\mathbf{\Lambda}_\beta^c := \sum_{i\in\mathcal{I}^c} \beta_i^c \mathbf{\Lambda}_i^c$, $\mathbf{\Lambda}_\beta^{\setminus c} := \sum_{i\in[N]} \beta_i^c \sum_{c'\in\mathcal{C}_i\setminus c} \mathbf{\Lambda}_i^{c'}$, where we recall that $\mathbf{\Lambda}_i^c = \mathbb{E}_{z\sim Q_{i,Z}^c}[zz^T]$. We further denote

$$\tilde{\mathbf{\Gamma}}_\beta^c := (\bar{\mathbf{\Lambda}}^c)^{\frac{1}{2}} (\mathbf{\Lambda}_\beta^c)^{-\frac{1}{2}}, \qquad \mathbf{\Gamma}_\beta^c := (\mathbf{\Lambda}_\beta^c)^{\frac{1}{2}} \left(\frac{\mathbf{\Lambda}_\beta^{\setminus c}}{m}\right)^{-\frac{1}{2}}.$$

In the above definitions, $\mathbf{\Lambda}_\beta^c$ is a linear combination of $\mathbf{\Lambda}_i^c$ for $i \in \mathcal{I}^c$, and a smaller $\|\tilde{\mathbf{\Gamma}}_\beta^c\|$ indicates that the $\mathbf{\Lambda}_\beta^c$ has better "coveragence" to $\bar{\mathbf{\Lambda}}^c$ in the feature space. Similar interpretations hold for $\mathbf{\Gamma}_\beta^c$.

We state our main theoretical results below and omit all detailed proofs due to space limitation.

Theorem 1: Define $\delta_{c,k} := \frac{10\sqrt{b_2 k}}{\sqrt{n^c}}$ and $\delta'_{c,k} := \frac{10\sqrt{b_3 k}}{\sqrt{n-n^c}}$ for some absolute constants b_2 and b_3, and define $a_\beta^c := \max_{i\in[N]} \beta_i^c/\hat{\beta}_i$. Suppose that Assumptions 1-4 hold and n^c, n are sufficiently large such that $\delta_{c,k} \leq 1/2$ and $\delta'_{c,k} \leq 1/2$. Then, for some absolute constant b, with probability at least $1-3e^{-100k}$, we have

$$\|w_c^* - \hat{w}_c\|_{\bar{\mathbf{\Lambda}}^c}^2 \leq 2\left(C_{\beta,c}^{\text{bias}}\Delta + \frac{LC_{\beta,c}^{\text{var}}}{m}\right)\big(1+\mathcal{O}\big(\delta_{c,k}+\delta'_{c,k}\big)\big), \tag{8}$$

where $L = \frac{b_3\sigma_\epsilon^2 k^{3/2}}{n}$, $C_{\beta,c}^{\text{var}} = a_\beta^c \big\|\mathbf{\Gamma}_\beta^c\big\|^4 \big\|\tilde{\mathbf{\Gamma}}_\beta^c\big\|^2$, and

$$C_{\beta,c}^{\text{bias}} = \frac{\big\|\tilde{\mathbf{\Gamma}}_\beta^c\big\|^2 \big\|(\mathbf{\Gamma}_\beta^c)^{-1}\big\|^2 \big\|\mathbf{\Gamma}_\beta^c\big\|^4}{\frac{1}{m^2}\big\|\mathbf{\Gamma}_\beta^c\big\|^4 + \frac{1}{m}\big\|\mathbf{\Gamma}_\beta^c\big\|^2 + 1}.$$

Remark 1: Note that L is the high-probability excess risk bound for k-dimensional linear regression with n IID samples and centered σ_ϵ^2 sub-gaussian noise, which scales in $\mathcal{O}(\frac{1}{n})$. The variance term in (8) scales in $\mathcal{O}(\frac{1}{nm})$, leading to an $\mathcal{O}\big(\frac{1}{m}\big)$ acceleration compared with the centralized IID setting. Meanwhile, the bias term contains a constant term $C_{\beta,c}^{\text{bias}}\Delta$ that does not diminish as n increases. This trade-off between variance and bias is due to the utilization of the augmented datasets. When Δ is sufficiently small, heterogeneous feature distributions can indeed help improve generalization.

Remark 2: To further gain some insight into the impact of weight assignment on this result, let us consider the case that $m \gg 1$. Then, $C_{\beta,c}^{\text{bias}} \approx \big\|\tilde{\mathbf{\Gamma}}_\beta^c\big\|^2 \big\|(\mathbf{\Gamma}_\beta^c)^{-1}\big\|^2 \big\|\mathbf{\Gamma}_\beta^c\big\|^4$, while $C_{\beta,c}^{\text{var}} = a_\beta^c \big\|\mathbf{\Gamma}_\beta^c\big\|^4 \big\|\tilde{\mathbf{\Gamma}}_\beta^c\big\|^2$. Although it is desirable to have $\big\|\mathbf{\Gamma}_\beta^c\big\|^2$, $\big\|\tilde{\mathbf{\Gamma}}_\beta^c\big\|$, and $\big\|(\mathbf{\Gamma}_\beta^c)^{-1}\big\|^2$ all as small as possible so that the excess risk bound can be reduced, we note that adjusting the assigned weight would affect them in a coupled fashion, leading to different bias-variance trade-offs.

B. Only inter-task heterogeneity

We next consider a slightly more restrictive case where only inter-task heterogeneity is considered. Specifically, we assume that for any task c, there exists a common distribution Q_z^c such that $Q_{i,Z}^c = Q_Z^c$ for any client i. We note that in general Q_Z^c may not be equal to $\bar{Q}_Z^c$. Besides, we also assume that all clients have the same amount of samples in their local datasets, i.e., $n_i = n_j$, $\forall i \neq j$. Define

$$\tilde{\mathbf{\Gamma}}^c := (\bar{\mathbf{\Lambda}}^c)^{\frac{1}{2}} (\mathbf{\Lambda}^c)^{-\frac{1}{2}}, \qquad \bar{\mathbf{\Gamma}}_\beta^c := (\mathbf{\Lambda}^c)^{\frac{1}{2}} \left(\frac{\mathbf{\Lambda}_\beta^{\setminus c}}{m\sum_{i\in\mathcal{I}^c}\beta_i^c}\right)^{-\frac{1}{2}},$$

where we denote $\mathbf{\Lambda}^c = \mathbb{E}_{z\sim Q_Z^c}[zz^T]$. Note that for a special case that $\bar{Q}_Z^c$ is a weighted combination of local feature distributions $Q_{i,Z}^c$ for $i \in \mathcal{I}^c$, in the inter-task heterogeneity only case, the matrix $\tilde{\mathbf{\Gamma}}^c$ is a k-dimensional identity matrix. Compared with the definitions of $\tilde{\mathbf{\Gamma}}_\beta^c$ and $\mathbf{\Gamma}_\beta^c$ in Section IV-A, $\tilde{\mathbf{\Gamma}}^c$ cannot be controlled via $\{\beta_i^c\}$ since intra-task heterogeneity no longer exists.

Before presenting the theoretical result, we introduce an additional assumption as follows.

Assumption 5: For any $z \sim Q_{i,Z}^{c'}$, $c' \neq c$, random variable $z^T w_c^*$ is centered, i.e., $\mathbb{E}\big[z^T w_c^*\big] = 0$.

Corollary 1: Suppose that Assumptions 1-5 hold. Then, for the inter-task heterogeneity only case, with probability at least $1-3e^{-100k}$ and $\delta_{c,k}$ and $\delta'_{c,k}$ defined in Theorem 1, we have

$$\|w_c^* - \hat{w}_c\|_{\bar{\mathbf{\Lambda}}^c}^2 \leq 2L\left(\frac{\Delta\tilde{C}_{\beta,c}^{\text{bias}}}{\sigma_\epsilon^2} + \frac{\tilde{C}_{\beta,c}^{\text{var}}}{m}\right)\big(1+\mathcal{O}\big(\delta_{c,k}+\delta'_{c,k}\big)\big),$$

where

$$\tilde{C}_{\beta,c}^{\text{bias}} = \frac{a_\beta^c}{\sum_{i\in\mathcal{I}^c}\beta_i^c} \cdot \frac{\big\|(\bar{\mathbf{\Gamma}}_\beta^c)^{-1}\big\|^2 \big\|\bar{\mathbf{\Gamma}}_\beta^c\big\|^4 \big\|\tilde{\mathbf{\Gamma}}^c\big\|^2}{\frac{1}{m^2}\big\|\bar{\mathbf{\Gamma}}_\beta^c\big\|^4 + \frac{1}{m}\big\|\bar{\mathbf{\Gamma}}_\beta^c\big\|^2 + 1},$$

$$\tilde{C}_{\beta,c}^{\text{var}} = \frac{a_\beta^c}{\sum_{i\in\mathcal{I}^c}\beta_i^c} \cdot \big\|\bar{\mathbf{\Gamma}}_\beta^c\big\|^4 \big\|\tilde{\mathbf{\Gamma}}^c\big\|^2.$$

Remark 3: Comparing with Theorem 1 where the bias term is a constant $2C_{\beta,c}^{\text{bias}}\Delta$, in Corollary 1 we show that with Assumption 5, the bias term becomes $\frac{2L\tilde{C}_{\beta,c}^{\text{var}}\Delta}{\sigma_\epsilon^2}$ and scales as $\mathcal{O}(\frac{\Delta}{\sigma_\epsilon^2 n})$. Essentially, with Assumption 5, the estimation error caused by the augmented dataset can be treated as a Δ-sub-gaussian noise. If $\frac{\Delta}{\sigma_\epsilon^2} \leq 1$, it implies that the added samples are more accurate to predict the task head w_c^* than the original samples in $(x,y) \in \mathcal{D}_i^c$.

Next, we investigate the excess risk under the model trained with the original datasets, and compare it with that using the augmented datasets. We have the following result.

Theorem 2: Suppose that Assumptions 1-4 hold. Let $\{\tilde{w}_c\}$ be the minimizer for (3). Then, with probability at least $1-2e^{-100k}$, the excess risk can be upper bounded as

$$\|\tilde{w}_c - w_c^*\|_{\bar{\mathbf{\Lambda}}^c}^2 \leq \frac{\tilde{a}_\beta^c}{\sum_{i\in\mathcal{I}^c}\beta_i^c} L^c \big\|\tilde{\mathbf{\Gamma}}^c\big\|^2 \big(1+\mathcal{O}(\delta_{c,k})\big), \tag{9}$$

where $L^c = \frac{b_3\sigma_\epsilon^2 k^{3/2}}{n^c}$.

Remark 4: Comparing Theorem 2 with Corollary 1, we see that using the augmented dataset $\tilde{\mathcal{D}}_i^c$ for FMTL leads to an excess risk upper bound that scales as $\mathcal{O}\left(\frac{\Delta\tilde{C}_{\beta,c}^{\text{bias}}}{\sigma_\epsilon^2} \cdot \frac{1}{n}\right)$, as opposed to $\mathcal{O}\left(\frac{\tilde{a}_\beta^c \|\tilde{\mathbf{\Gamma}}^c\|^2}{\sum_i \beta_i^c} \cdot \frac{1}{n^c}\right)$ when $\mathcal{D}_i^c$ is used. Hence, as

long as $\frac{\Delta\tilde{C}^{\text{bias}}_{\beta,c}}{\sigma^2_\epsilon}\cdot\frac{1}{n}\ll\frac{\tilde{a}^c_\beta\|\tilde{\mathbf{\Gamma}}^c\|^2}{\sum_i\beta^c_i}\cdot\frac{1}{n^c}$ (e.g., with large M such that $n\gg n^c$ for any c), using $\tilde{\mathcal{D}}^c_i$ for linear regression helps improve the generalization performance.

V. Excess risk-Aware Weight assignment

Corollary 1 indicates that the bias term would dominate the excess risk when $m\gg 1$. Also note that $\tilde{C}^{\text{bias}}_{\beta,c}\leq C\frac{a^{\max}_\beta}{\sigma_{\min}\left(\mathbf{\Lambda}^{\setminus c}_\beta\right)\sum_{i\in\mathcal{I}^c}\beta^c_i}$, where C is a constant not related to $\{\beta^c_i\}$. Thus, to reduce the excess risk, for any task c, we can choose $\{\beta^c_i\}$ by solving the following optimization problem:

$$\begin{aligned}\min_{\{\beta^c_i\}_{i\in[N]}}\ &\frac{a^c_\beta}{\sigma_{\min}\left(\mathbf{\Lambda}^{\setminus c}_\beta\right)\sum_{i\in\mathcal{I}^c}\beta^c_i}\\ \text{s.t.}\ \ &\sum_{i=1}^N\beta^c_i=1;\ \beta^c_i\geq 0,\forall i\in[N].\end{aligned}\tag{10}$$

Note that problem (10) is generally non-convex and $\mathbf{\Lambda}^{\setminus c}_\beta$ may be difficult to estimate in practice. As a first step, we consider a special case where $m=1$, $n_i=n/N$ for all clients, and there exists a constant $\tilde{b}_3$ such that $\left\|(\mathbf{\Lambda}^c)^{-1}\right\|\leq\tilde{b}_3$ almost surely for any c.

Now we consider to assign weights for task c_1 at all clients, i.e., $\{\beta^{c_1}_i\}_{i\in[N]}$. We can show that

$$\begin{aligned}\tilde{C}^{\text{bias}}_{\beta,c_1}&\leq\tilde{b}_1\tilde{b}_2\frac{a^{c_1}_\beta}{\sum_{i\in\mathcal{I}^{c_1}}\beta^{c_1}_i}\left\|(\mathbf{\Lambda}^{\setminus c_1}_\beta)^{-1}\right\|\\&\leq\tilde{b}_1\tilde{b}_2\tilde{b}_3\frac{a^{c_1}_\beta}{(\sum_{i\in\mathcal{I}^{c_1}}\beta^{c_1}_i)^2(\sum_{i\notin\mathcal{I}^{c_1}}\beta^{c_1}_i)},\end{aligned}\tag{11}$$

where the last inequality holds since $\sum_{i\in\mathcal{I}^{c_1}}\beta^{c_1}_i\leq 1$.

Based on the definition of a^c_β, we can show that the right hand side of (11) can be minimized by setting

$$\beta^{c_1}_i=\begin{cases}\frac{1}{2|\mathcal{I}^{c_1}|}, & \text{if } i\in\mathcal{I}^{c_1}\text{ and }|\mathcal{I}^{c_1}|\leq\frac{N}{2}\\ \frac{1}{2\left(N-|\mathcal{I}^{c_1}|\right)}, & \text{if } i\notin\mathcal{I}^{c_1}\text{ and }|\mathcal{I}^{c}|\leq\frac{N}{2}\\ \frac{1}{N}, & \text{if } |\mathcal{I}^{c_1}|>\frac{N}{2}.\end{cases}\tag{12}$$

With the weight selection method in (12), we can modify the well-known FedAvg algorithm [1] for improved generalization performance for multi-task FL. Moreover, if a good representation ϕ is not known beforehand, we develop a two-phased **FedAvg+FMTL** algorithm that learns the common representation using FedAvg across all tasks in the first phase (e.g., the first T_1 communication rounds), and then trains the class-specific layers based on the learned representation layers in the second phase (e.g., the remaining $T-T_1$ communication rounds). The complete FedAvg+FMTL algorithm is omitted due to space limitation.

VI. Experimental results

In this section, we evaluate the performance of FMTL using real-world datasets CIFAR-10 and CIFAR-100. For CIFAR-10, we consider three (N,m) pairs: $(100,2)$, $(100,5)$, and $(1000,2)$. We train a convolutional neural network (CNN) with two convolution layers, two fully connected layers with ReLU activation, and a final fully connected layer with a softmax activation function. For the final 64×10 layer, we set its c-th column as a task-specific head w_c for task c. For CIFAR-100, we use the following three (N,m) pairs: $(100,5)$, $(100,10)$, and $(100,20)$. In this case, we train a CNN model with two convolution layers followed by three fully connected layers with ReLU activation, and the last fully connected layer is followed by a softmax activation function. Additionally, a dropout layer is incorporated between the two convolutional layers. In this case, we set its c-th column of the final 128×100 layer as a task-specific head w_c for task c.

We run the proposed **FedAvg+FMTL** algorithm with the weight assignment in (12) for 110 epochs with $T_1=100$. For comparison, we also run the vanilla **FedAvg** with 110 epochs, and for **FedAvg+EW** we run FedAvg for 100 epochs and only train the output layer for 10 epochs with the empirical weights $\beta_i=\frac{1}{|\mathcal{I}^i|}$. For FedProx [29], three experiments are conducted: (1) **FedProx+FMTL** simply changes the algorithm for training the representation layer in FedAvg+FMTL from FedAvg to FedProx, and we run FedProx+FMTL for 110 epochs with weights given in (12). (2) As for **FedProx+EW**, we train FedProx for 100 epochs and then only train on the output layer for another 10 epochs with empirical weights. (3) As a baseline of comparison, we also train FedProx for 100 epochs. The results are summarized in Table I. We can see that **FedAvg/Fedprox+FMTL** has the best test accuracy on CIFAR-10 and CIFAR-100 among all the methods.

TABLE I
Average test accuracy on CIFAR-10 and CIFAR-100

(# clients× # tasks/client)		100× 2	100 × 5	1000× 2
CIFAR-10	FedAvg	40.417	51.940	39.856
	FedAvg+FMTL	**45.163**	**54.449**	**43.127**
	FedAvg+EW	41.112	51.215	39.084
	FedProx	40.788	52.904	34.020
	FedProx+FMTL	**45.965**	**54.322**	**34.730**
	FedProx+EW	41.475	50.934	33.510
(# clients× # tasks/client)		100× 5	100 × 10	100× 20
CIFAR-100	FedAvg	22.397	25.110	30.017
	FedAvg+FMTL	**25.934**	**28.362**	**32.564**
	FedAvg+EW	22.250	25.804	30.347
	FedProx	21.285	24.758	29.486
	FedProx+FMTL	**24.742**	**26.602**	**30.260**
	FedProx+EW	20.318	23.860	28.368

VII. Conclusions

In this work, we studied the impact of feature heterogeneity on the generalization performance of FMTL. We analyzed the excess risk upper bound for a federated multi-task linear regression problem. Under certain feature heterogeneity assumptions, we proposed a new augmented dataset based approach, and proved that the excess risk bound can be improved by a factor up to $\mathcal{O}(1/m)$ compared with that under a centralized homogeneous setting, where m is the number of tasks performed at each client. A simple weight assignment method based on minimizing the excess risk upper bound was then proposed, whose effectiveness was validated through numerical experiments.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *AISTATS*. PMLR, 2017, pp. 1273–1282.

[2] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[3] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5132–5143.

[4] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "Fedbn: Federated learning on non-iid features via local batch normalization," *arXiv preprint arXiv:2102.07623*, 2021.

[5] R. Pathak and M. J. Wainwright, "Fedsplit: An algorithmic framework for fast federated optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7057–7066, 2020.

[6] M. Al-Shedivat, J. Gillenwater, E. Xing, and A. Rostamizadeh, "Federated learning via posterior averaging: A new perspective and practical algorithms," *arXiv preprint arXiv:2010.05273*, 2020.

[7] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4615–4625.

[8] H. Yuan, W. Morningstar, L. Ning, and K. Singhal, "What do we mean by generalization in federated learning?" *arXiv preprint arXiv:2110.14216*, 2021.

[9] H. Yuan, W. R. Morningstar, L. Ning, and K. Singhal, "What do we mean by generalization in federated learning?" in *International Conference on Learning Representations*, 2022.

[10] B. Bullins, E. Hazan, A. Kalai, and R. Livni, "Generalize across tasks: Efficient algorithms for linear representation learning," in *Algorithmic Learning Theory*. PMLR, 2019, pp. 235–246.

[11] S. S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei, "Few-shot learning via learning the representation, provably," *arXiv preprint arXiv:2002.09434*, 2020.

[12] N. Tripuraneni, C. Jin, and M. Jordan, "Provable meta-learning of linear representations," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 434–10 443.

[13] A. Maurer, M. Pontil, and B. Romera-Paredes, "The benefit of multitask representation learning," *Journal of Machine Learning Research*, vol. 17, no. 81, pp. 1–32, 2016.

[14] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[15] M. G. Arivazhagan, V. Aggarwal, A. K. Singh, and S. Choudhary, "Federated learning with personalization layers," *arXiv preprint arXiv:1912.00818*, 2019.

[16] P. P. Liang, T. Liu, L. Ziyin, N. B. Allen, R. P. Auerbach, D. Brent, R. Salakhutdinov, and L.-P. Morency, "Think locally, act globally: Federated learning with local and global representations," *arXiv preprint arXiv:2001.01523*, 2020.

[17] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Exploiting shared representations for personalized federated learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 2089–2099.

[18] A. Zhong, H. He, Z. Ren, N. Li, and Q. Li, "Feddar: Federated domain-aware representation learning," *arXiv preprint arXiv:2209.04007*, 2022.

[19] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Fedavg with fine tuning: Local updates lead to representation learning," *arXiv preprint arXiv:2205.13692*, 2022.

[20] A. Rakotomamonjy, M. Vono, H. J. M. Ruiz, and L. Ralaivola, "Personalised federated learning on heterogeneous feature spaces," *arXiv preprint arXiv:2301.11447*, 2023.

[21] T. Yu, E. Bagdasaryan, and V. Shmatikov, "Salvaging federated learning by local adaptation," *arXiv preprint arXiv:2002.04758*, 2020.

[22] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *International Conference on Machine Learning*. PMLR, 2021, pp. 6357–6368.

[23] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards personalized federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

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

[25] L. Deng, "The MNIST database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[26] H. Kusetogullari, A. Yavariabdi, J. Hall, and N. Lavesson, "Digitnet: a deep handwritten digit detection and recognition method using a new historical handwritten digit dataset," *Big Data Research*, vol. 23, p. 100182, 2021.

[27] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," 2011.

[28] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint arXiv:1011.3027*, 2010.

[29] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *The 3rd MLSys Conference*, 2020.

[30] D. Hsu, S. M. Kakade, and T. Zhang, "Random design analysis of ridge regression," in *Conference on learning theory*. JMLR Workshop and Conference Proceedings, 2012, pp. 9–1.