On Sample Complexity of Learning Shared Representations: The Asymptotic Regime

Xinyi Tong[1], Xiangxiang Xu[2], and Shao-Lun Huang[3]

***Abstract*— Federated learning and multi-task learning have achieved great successes in recent works, which exploit the knowledge contained in different tasks to alleviate the influence of lacking training samples. In order to understand the knowledge utilization mechanism, we study the sample complexity of the federated learning algorithms by computing their expected population risks, where the asymptotic regime is considered that the number of tasks is large. Specifically, we focus on several typical algorithms, including (i) training personalized models merely by local data, (ii) jointly training a single model by all the data of devices, and (iii) jointly training the shared feature representation and learning individual classifiers of different devices, where the latter one is recently widely-adopted. The results reveal that the validity of the shared representation algorithm is affected by the task similarities, the model dimensionality, and the sample size, which quantify the sample size range where such algorithm is preferable. Moreover, we develop a new algorithm in consideration of collaborating the classifiers of different devices under the shared representations, which can have lower sample complexity than conventional methods under the asymptotic regime of large sample size. This provides theoretical guidance for practical algorithm designs.**

I. Introduction

Simultaneously solving multiple machine learning tasks on the same domain of data has gained much attention in federated learning and multi-task learning [1]. For example, it is often considered in federated learning that there are many distributed devices and a server, where each device can collect data for solving its own learning task. The learning tasks across the devices can be correlated, and the devices can interact with the server for jointly solving the learning tasks. In multi-task learning, there are several correlated machine learning tasks with different training data, and the goal is to jointly solve these tasks by using the task correlations and available data. One critical challenge for the abovementioned problems is that the training data size of each task may not be sufficient for training a good machine learning model. As such, the interactions between different tasks and certain kind of data sharing in joint model training are often necessary for achieving better learning performance. A naive approach considered in federated learning for addressing this problem is to learning a single model for all different tasks via the training data for all tasks. When the learning tasks across the devices are considerably similar, and the size of the training data from all devices are sufficiently large, the overall performance of all the learning tasks can be improved comparing to solving each task with its own data. However, when the learning tasks of different devices are not quite similar, such approach can lead to significant performance degradation [2]. Therefore, instead of sharing the entire machine learning model across different tasks, it is recently proposed in [3] to share only the feature representation of the data with different classifiers. Specifically, this approach applies all training data to train a shared feature representation for all tasks, and then each task trains its own linear classifier according to the shared feature representation and its own training data. Intuitively, such approach can perform well when different tasks are correlated but not considerably similar, and has gained successes in some applications. Nevertheless, the theoretical understanding of why and when learning shared representation can be beneficial appears to be quite behind. Therefore, our goal is to provide a theoretical framework to quantify the positive/negative gains of learning shared representation in different scenarios.

In this paper, we consider the asymptotic regime, such that the number of tasks k is large, which is practical in many federated learning applications [4]. Under such regime, we first investigate three typical approaches in federated learning: (i) Each device learns a model only by its own training data; (ii) All devices jointly learn a single model for all tasks by all training data; (iii) All devices jointly learn a shared representation by all training data, and each device learns its own classifier by its training data. These three approaches are illustrated in Figure 1. The learning model of these approaches are trained by the empirical risk minimization (ERM) through the training data, and the performance of the algorithms are evaluated by the expected population risks (EPR). In particular, we demonstrate analytical characterizations of the EPRs in the asymptotic regime for the three approaches in terms of the underlying distributions and the number of training samples of each task. Then, we compare the EPRs for different approaches, and specify the parameter regimes such that learning shared representation can be beneficial comparing to other approaches. Such regime can be expressed as a comparison between the task similarities between devices and the number of training samples at each device, which quantifies the intuition of learning shared representations.

Moreover, it is interesting to note that the approach (iii) does not exploit the correlations between the classifiers. Hence, we extend the approach (iii) by allowing the classifiers to collaborate via linear filtering. Thus, the classifier at

[1]Xinyi Tong is with Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Shenzhen 518055, China txy18@mails.tsinghua.edu.cn

[2]Xiangxiang Xu is with EECS, MIT, Cambridge, MA 02139, USA xuxx@mit.edu

[3]Shao-Lun Huang is with the Data Science and Information Technology Research Center, Tsinghua-Berkeley Shenzhen Institute, Shenzhen 518055, China shaolun.huang@sz.tsinghua.edu.cn

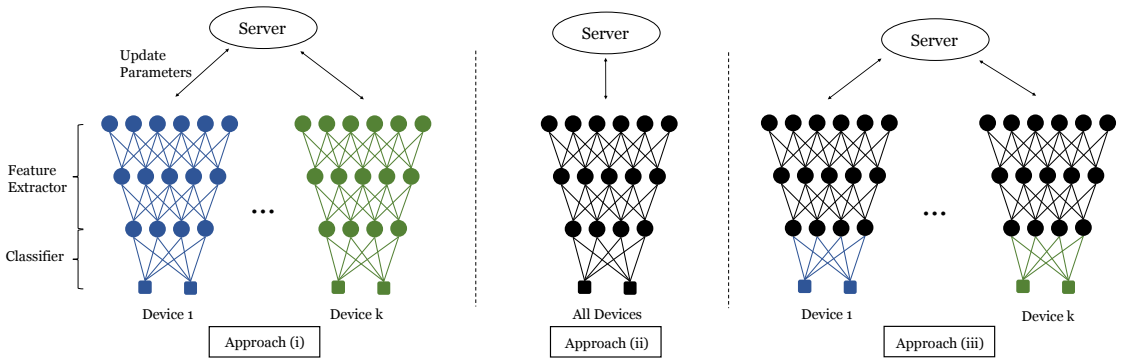

Fig. 1. A diagram to show the three federated learning algorithms, where the learning model can be separated as the feature extractor with many parameters and the classifier with few ones. Particularly, Approach (i) learns an individual model of each device only by its own data; Approach (ii) jointly learn a single model for all tasks; and Approach (iii) learns a shared feature extractor by all training data, where each device learns its own classifier.

each device in this new approach is represented by linear combinations of the classifiers from the approach (iii), where the combining coefficients are subjected to be designed to optimize the EPR. We present the analytical solution of the optimal coefficients in terms of the underlying distributions of the tasks and the number of training samples. This leads to a lower bound for the EPR of this new approach, since such optimal coefficients are not computable without the knowledge of the underlying distributions. Then, we propose a computable algorithm which estimates the coefficients by the training data, and present theoretical analyses of the EPR performance of this algorithm. Our results demonstrate the performance gain of the classifier collaborations in terms of the task similarities and the number of samples in the asymptotic regime. This provides the theoretical guidance for algorithm designs in federated learning and multi-task learning problems.

The contributions of this paper are summarized as follows:

1) We provide the sample complexity analyses of the aforementioned typical federated algorithms under the discrete data space by computing their EPRs under the large task number regime. Based on the results, we confirm the sample size range that the shared representation algorithm could outperform the other two algorithms and interpret the influence of the task similarity and the model dimensionality.
2) We improve the shared representation algorithm by collaborating the classifiers learned in different devices, where a lower bound of its EPR can be given under the optimal collaboration. We also provide a practical collaboration approach and verify its effectiveness under the regime of large sample sizes.
3) We extend the sample complexity analyses to the continuous sample space and provide the EPRs of different algorithms under the large sample size regime with respect to the number of parameters contained in the learning models, where similar interpretations as in the discrete case are provided.

A. Related Works

In consideration of the fact that approaches (i) and (ii) are not particularly intended for federated learning [5] and multi-task learning [6], we maintain our main focus on approach (iii). Such a shared representation method is the major part of the representation learning area, which focuses on building up joint representations among different tasks [1]. The successes of shared representation methods have been demonstrated when multiple tasks are simultaneously tackled, e.g., in the image classification tasks [7]. Finally, theoretical analyses in this area are in progress recently, which mainly cover the typical Gaussian signal analyses based on the ERM principle [8], and the mechanism of concrete methods, e.g., low-rank method and task clustering method [9]. Compared with previous works, this paper can provide insights of how the sample sizes affect the effectiveness of the shared representation method based on a more general framework. On the other hand, since federated learning is concentrated on privacy and computation capacity, theoretical analyses in this area mainly focus on the convergence rate from the perspective of optimization [10].

II. Problem Formulation

A. Notations

Let X and Y be the random variables of the data and label with domains $\mathcal{X}$ and $\mathcal{Y}$, respectively. For ease of illustration, we assume X to be discrete in Section II and Section III, and will extend our analyses to continuous cases later.

To start with, we consider that the federated learning problem has k tasks. For each task of device $i = 1, \cdots, k$, we assume that n training samples $\{(x_\ell^{(i)}, y_\ell^{(i)})\}_{\ell=1}^n$ are i.i.d. generated from some underlying joint distribution $P_{XY}^{(i)}$, and the corresponding empirical distribution of samples is defined as $\hat{P}_{XY}^{(i)}(x,y) \triangleq \frac{1}{n}\sum_{l=1}^n \mathbb{1}\left\{x_l^{(i)} = x, y_l^{(i)} = y\right\}$, where $\mathbb{1}\{\cdot\}$ denotes the indicator function [11].

In this paper, we study the sample complexity of federated learning algorithms under the asymptotic regime, where we make the following assumption:

Assumption 2.1 (Asymptotic Regime Assumption): The task number k is arbitrarily large.

To describe the asymptotic behaviors of the related EPRs when $k \to \infty$, we employ the big O notations [12], which is with respect to k for the whole paper. Specifically, when computing the EPR of some algorithms, we also assume the sample size n to be large [cf. Assumption 3.6]. Under this circumstance, the big O notation is also with respect to n.

In particular, we assume that all the joint distributions $P_{XY}^{(i)}$'s share the same marginal distributions of data and label, denoted as P_X[1] and P_Y, which is natural with all devices sharing the same domains and widely-adopted in practical federated learning implementations of i.i.d. data case [13]. We also make a mild assumption that each label occurs with the same probability, i.e., $P_Y(y) = 1/|\mathcal{Y}|$ for each $y \in \mathcal{Y}$, which is satisfied in most of the popular experimental settings [14]. Moreover, the underlying distribution of each device i is assumed to be decomposed by the feature $\boldsymbol{f}_i : \mathcal{X} \to \mathbb{R}^d$ and the classifier $\boldsymbol{g}_i : \mathcal{Y} \to \mathbb{R}^d$, where $d \geq |\mathcal{Y}|$, i.e., the (x, y)th entry of $P_{XY}^{(i)}$ is

$$P_{XY}^{(i)}(x, y) = P_X(x)P_Y(y)\left(1 + \boldsymbol{f}_i^{\mathrm{T}}(x)\boldsymbol{g}_i(y)\right). \quad (1)$$

Furthermore, the feature and classifier are required to be zero-mean under the marginal distributions[2], i.e., $\mathbb{E}_{P_X}[\boldsymbol{f}_i(X)] = \mathbb{E}_{P_Y}[\boldsymbol{g}_i(X)] = \mathbf{0}$. Then, we define the *canonical dependence matrix* (CDM) [15] $\mathbf{B}(P_{XY}) \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ of any distribution P_{XY} supported by $\mathcal{X} \times \mathcal{Y}$:

$$[\mathbf{B}(P_{XY})](y, x) \triangleq \frac{P_{XY}(x, y) - P_X(x)P_Y(y)}{\sqrt{P_X(x)P_Y(y)}}, \quad (2)$$

where $[\mathbf{B}(P_{XY})](y, x)$ is the (y, x)th entry of $\mathbf{B}(P_{XY})$.

We also define the *feature matrix* $\boldsymbol{\Phi}(\boldsymbol{f}) \in \mathbb{R}^{|\mathcal{X}|\times d}$ of any feature function $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^d$:

$$\boldsymbol{\Phi}(\boldsymbol{f}) \triangleq \left[\sqrt{P_X(1)}\boldsymbol{f}^{\mathrm{T}}(1), \cdots, \sqrt{P_X(|\mathcal{X}|)}\boldsymbol{f}^{\mathrm{T}}(|\mathcal{X}|)\right]^{\mathrm{T}}, \quad (3)$$

and the *classifier matrix* $\boldsymbol{\Psi}(\boldsymbol{g}) \in \mathbb{R}^{|\mathcal{Y}|\times d}$ of any label classifier $\boldsymbol{g} : \mathcal{Y} \to \mathbb{R}^d$:

$$\boldsymbol{\Psi}(\boldsymbol{g}) \triangleq \left[\sqrt{P_Y(1)}\boldsymbol{g}^{\mathrm{T}}(1), \cdots, \sqrt{P_Y(|\mathcal{Y}|)}\boldsymbol{g}^{\mathrm{T}}(|\mathcal{Y}|)\right]^{\mathrm{T}}. \quad (4)$$

For simplicity, we define the corresponding matrices of each device i as $\boldsymbol{\Phi}_i \triangleq \boldsymbol{\Phi}(\boldsymbol{f}_i)$, $\boldsymbol{\Psi}_i \triangleq \boldsymbol{\Psi}(\boldsymbol{g}_i)$, $\forall i = 1, \cdots, k$. With such notations, the CDM of the underlying distribution of each device i is $\mathbf{B}(P_{XY}^{(i)}) = \boldsymbol{\Psi}_i\boldsymbol{\Phi}_i^{\mathrm{T}}$. Notice the fact that $\mathbf{B}(P_{XY}^{(i)})$ keeps invariant under the invertible linear mapping in the feature space $\mathbb{R}^d$, i.e., for any invertible matrix $\mathbf{Q} \in \mathbb{R}^{d\times d}$, $\boldsymbol{\Psi}_i\boldsymbol{\Phi}_i^{\mathrm{T}} = \boldsymbol{\Psi}_i\mathbf{Q}^{-1}\mathbf{Q}\boldsymbol{\Phi}_i^{\mathrm{T}}$. We can without loss of generality assume that for each $i = 1, \cdots, k$, $\boldsymbol{\Phi}_i^{\mathrm{T}}\boldsymbol{\Phi}_i = \mathbf{I}_d$, where $\mathbf{I}_d$ denotes the identity matrix in $\mathbb{R}^{d\times d}$.

Finally, we introduce the Frobenius matrix norm [16]: for each matrix $\mathbf{A} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$, $\|\mathbf{A}\|_{\mathrm{F}}^2 \triangleq \mathrm{tr}\left(\mathbf{A}\mathbf{A}^{\mathrm{T}}\right) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}}[\mathbf{A}(x, y)]^2$. Such norm is employed for evaluating the performance of the learned features and classifiers. In particular, when $\hat{\boldsymbol{f}} : \mathcal{X} \to \mathbb{R}^d$ and $\hat{\boldsymbol{g}} : \mathcal{Y} \to \mathbb{R}^d$ are learned from samples for device i, the expected population risk of the learned features[3] is

$$\begin{aligned} R_i\left(\hat{\boldsymbol{f}}, \hat{\boldsymbol{g}}\right) &\triangleq \mathbb{E}\left[\left\|\mathbf{B}\left(P_{XY}^{(i)}\right) - \boldsymbol{\Psi}(\hat{\boldsymbol{g}})\boldsymbol{\Phi}^{\mathrm{T}}(\hat{\boldsymbol{f}})\right\|_{\mathrm{F}}^2\right] \\ &= \mathbb{E}\left[\left\|\boldsymbol{\Psi}_i\boldsymbol{\Phi}_i^{\mathrm{T}} - \boldsymbol{\Psi}(\hat{\boldsymbol{g}})\boldsymbol{\Phi}^{\mathrm{T}}(\hat{\boldsymbol{f}})\right\|_{\mathrm{F}}^2\right], \end{aligned} \quad (5)$$

where the expectation is taken over all possible empirical distributions $\hat{P}_{XY}^{(i)}$ following the multinomial distributions [19].

B. Federated learning algorithms

First, we evaluate three typical federated learning algorithms as defined in the following.

Personalized Model Algorithm (Algorithm (i)):

When the samples of each device are sufficient to learn a well-performed model, devices do not require assistance from the others. Under this circumstance, the most trivial way is to learn the personalized model by its own training data, which means for $i = 1, \cdots, k$, the feature of data and the classifier of label can be achieved by minimizing the empirical risk

$$\begin{aligned} &\left(\hat{\boldsymbol{f}}_{\mathtt{PM}i}, \hat{\boldsymbol{g}}_{\mathtt{PM}i}\right) \\ &= \underset{(\boldsymbol{f},\boldsymbol{g}):\boldsymbol{\Psi}^{\mathrm{T}}(\boldsymbol{f})\boldsymbol{\Psi}(\boldsymbol{f})=\mathbf{I}_d}{\arg\min} \left\|\mathbf{B}\left(\hat{P}_{XY}^{(i)}\right) - \boldsymbol{\Psi}(\boldsymbol{g})\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{f})\right\|_{\mathrm{F}}^2. \end{aligned} \quad (6)$$

The corresponding EPR is

$$R_{\mathtt{PM}} \triangleq \sum_{i=1}^{k} R_i\left(\hat{\boldsymbol{f}}_{\mathtt{PM}i}, \hat{\boldsymbol{g}}_{\mathtt{PM}i}\right). \quad (7)$$

Global Model Algorithm (Algorithm (ii)):

When local samples are extremely insufficient, all devices are forced to learn a global model jointly, otherwise the random noise from the sampling process can lead to a large generalization error. Under this circumstance, the global feature of data and classifier of label can be achieved by

$$\begin{aligned} &\left(\hat{\boldsymbol{f}}_{\mathtt{GM}}, \hat{\boldsymbol{g}}_{\mathtt{GM}}\right) \\ &= \underset{(\boldsymbol{f},\boldsymbol{g}):\boldsymbol{\Psi}^{\mathrm{T}}(\boldsymbol{f})\boldsymbol{\Psi}(\boldsymbol{f})=\mathbf{I}_d}{\arg\min} \left\|\mathbf{B}\left(\frac{1}{k}\sum_{i=1}^{k}\hat{P}_{XY}^{(i)}\right) - \boldsymbol{\Psi}(\boldsymbol{g})\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{f})\right\|_{\mathrm{F}}^2. \end{aligned} \quad (8)$$

The corresponding EPR is

$$R_{\mathtt{GM}} \triangleq \sum_{i=1}^{k} R_i\left(\hat{\boldsymbol{f}}_{\mathtt{GM}}, \hat{\boldsymbol{g}}_{\mathtt{GM}}\right). \quad (9)$$

Shared Representation Algorithm (Algorithm (iii)):

As introduced in [3], this typical federated learning algorithm shares only the feature representation of the data and learns different classifiers. Under this circumstance, the joint feature representation and individual classifiers of different

[1]Without loss of generality, we assume all the entries $P_X(x) > 0$, or we can exclude this element x from the space $\mathcal{X}$.

[2]This is not an extra assumption, which is a conclusion of the canonical correlation analysis of joint distribution decomposition. [cf. Proposition 5 [15]]

[3]Such performance measure based on CDM can be recognized as a good approximation of Kullback–Leibler (K-L) divergence [17] between the underlying distribution $P_{XY}^{(i)}(x, y)$ and the distribution model with learned features $Q_{XY}(x, y) \triangleq P_X(x)P_Y(y)(1 + \hat{\boldsymbol{f}}^{\mathrm{T}}(x)\hat{\boldsymbol{g}}(y))$, where $D(P_{XY}^{(i)}\|Q_{XY}) \approx \frac{1}{2}\left\|\mathbf{B}(P_{XY}^{(i)}) - \boldsymbol{\Psi}(\hat{\boldsymbol{g}})\boldsymbol{\Phi}^{\mathrm{T}}(\hat{\boldsymbol{f}})\right\|_{\mathrm{F}}^2$. [18]

devices can be achieved by an iterative algorithm, whose results satisfy

$$\hat{\boldsymbol{f}}_{\texttt{SR}} = \underset{\boldsymbol{f}:\boldsymbol{\Psi}^{\mathrm{T}}(\boldsymbol{f})\boldsymbol{\Psi}(\boldsymbol{f})=\mathbf{I}_d}{\arg\min} \sum_{i=1}^{k} \left\| \mathbf{B}\left(\hat{P}_{XY}^{(i)}\right) - \boldsymbol{\Psi}(\hat{\boldsymbol{g}}_{\texttt{SR}i})\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{f}) \right\|_{\mathrm{F}}^2, \tag{10}$$

and

$$\hat{\boldsymbol{g}}_{\texttt{SR}i} = \underset{\boldsymbol{g}}{\arg\min} \left\| \mathbf{B}\left(\hat{P}_{XY}^{(i)}\right) - \boldsymbol{\Psi}(\boldsymbol{g})\boldsymbol{\Phi}^{\mathrm{T}}(\hat{\boldsymbol{f}}_{\texttt{SR}}) \right\|_{\mathrm{F}}^2. \tag{11}$$

The corresponding EPR is

$$R_{\texttt{SR}} \triangleq \sum_{i=1}^{k} R_i\left(\hat{\boldsymbol{f}}_{\texttt{SR}}, \hat{\boldsymbol{g}}_{\texttt{SR}i}\right). \tag{12}$$

Classifiers Collaboration Algorithm (Algorithm (iv)):

As illustrated in Section I, Algorithms (i)-(iii) are the classical algorithms that have been widely-applied in practical implementations. Moreover, we consider that the correlations between the classifiers are not fully exploited, where the classifiers learned in Algorithm (iii) can be collaborated via linear filtering. We simply consider the new classifier of device i as a linear combination of the classifier $\hat{\boldsymbol{g}}_{\texttt{SR}i}$ trained by its own data and the averaged learned classifier $\hat{\boldsymbol{g}}_{\texttt{SR}}$, where $\hat{\boldsymbol{g}}_{\texttt{SR}} \triangleq \frac{1}{k}\sum_{i=1}^{k} \hat{\boldsymbol{g}}_{\texttt{SR}i}$. Then, we have the following classifier for device i:

$$\hat{\boldsymbol{g}}_{\texttt{CC}i} = \alpha_i \hat{\boldsymbol{g}}_{\texttt{SR}i} + (1-\alpha_i)\hat{\boldsymbol{g}}_{\texttt{SR}}, \tag{13}$$

where α_i is a function of all the data as given in Section III-B. The corresponding EPR is

$$R_{\texttt{CC}} \triangleq \sum_{i=1}^{k} R_{\texttt{CC}i}, \text{ where } R_{\texttt{CC}i} \triangleq R_i\left(\hat{\boldsymbol{f}}_{\texttt{SR}}, \hat{\boldsymbol{g}}_{\texttt{CC}i}\right). \tag{14}$$

III. EPR Analyses: The Discrete Case

A. Analyses for classical algorithms

First, we directly establish the analytical expressions of the EPRs (7), (9), and (12). The proofs are provided in the supplementary material.

1) Personalized model algorithm: For the Personalized Model Algorithm (Algorithm (i)), it can be easily derived from the definition (6) that the learned $\hat{\boldsymbol{f}}_{\texttt{PM}i}$ and $\hat{\boldsymbol{g}}_{\texttt{PM}i}$ correspond to the empirical distribution $\hat{P}_{XY}^{(i)}$ as the objective function being zero, i.e., $\boldsymbol{\Psi}(\hat{\boldsymbol{g}}_{\texttt{PM}i})\boldsymbol{\Phi}^{\mathrm{T}}(\hat{\boldsymbol{f}}_{\texttt{PM}i}) = \mathbf{B}(\hat{P}_{XY}^{(i)})$. Then, we have the following characterization of the EPR (7).

Theorem 3.1 (EPR of Personalized Model Algorithm): The expected population risk as defined in (7) is

$$R_{\texttt{PM}} = \frac{k|\mathcal{X}||\mathcal{Y}| - k + \sum_{i=1}^{k} C_i}{n}, \tag{15}$$

where $C_i \triangleq \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \boldsymbol{f}_i^{\mathrm{T}}(x)\boldsymbol{g}_i(y) - \|\boldsymbol{\Psi}_i\|_{\mathrm{F}}^2$, and $\boldsymbol{\Psi}_i$ is the classifier matrix of $\boldsymbol{g}_i$.

Let $C_{\min} \triangleq \min\limits_{x\in\mathcal{X},y\in\mathcal{Y},i\in\{1,\cdots,k\}} \{\boldsymbol{f}_i^{\mathrm{T}}(x)\boldsymbol{g}_i(y)\} - \frac{1+|\mathcal{Y}|}{|\mathcal{X}||\mathcal{Y}|}$ and $C_{\max} \triangleq \max\limits_{x\in\mathcal{X},y\in\mathcal{Y},i\in\{1,\cdots,k\}} \{\boldsymbol{f}_i^{\mathrm{T}}(x)\boldsymbol{g}_i(y)\}$. Then, $R_{\texttt{PM}}$ can be bounded[4] by

$$\frac{k\left(1+C_{\min}\right)|\mathcal{X}||\mathcal{Y}|}{n} \le R_{\texttt{PM}} \le \frac{k\left(1+C_{\max}\right)|\mathcal{X}||\mathcal{Y}|}{n}.$$

Additionally, Theorem 3.1 states that the Personalized Model Algorithm provides an unbiased estimation of the underlying distributions $P_{XY}^{(i)}$'s and all the EPRs come from the variances of the sampling process, which is inversely proportional to the sample size n of each device. When estimating the underlying distributions, there are $|\mathcal{X}||\mathcal{Y}|$ parameters (number of entries) contained in the distribution models, which appear in the EPR (15) and reflect the model dimensionality.

2) Global Model Algorithm: As for the Global Model Algorithm (Algorithm (ii)), the feature and classifier are learned with nk samples of all the devices, which lead to an extremely small variance and the main EPR comes from the bias by using the global empirical distribution $\frac{1}{k}\sum_{i=1}^{k}\hat{P}_{XY}^{(i)}$ to estimate each device's distribution. Let

$$(\boldsymbol{f}_{\texttt{GM}}, \boldsymbol{g}_{\texttt{GM}}) \triangleq \underset{(\boldsymbol{f},\boldsymbol{g}):\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{f})\boldsymbol{\Phi}(\boldsymbol{f})=\mathbf{I}_d}{\arg\min} \left\| \mathbf{B}\left(\frac{1}{k}\sum_{i=1}^{k} P_{XY}^{(i)}\right) - \boldsymbol{\Psi}(\boldsymbol{g})\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{f}) \right\|_{\mathrm{F}}^2. \tag{16}$$

It is easy to verify that $\boldsymbol{\Psi}(\boldsymbol{g}_{\texttt{GM}})\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{f}_{\texttt{GM}}) = \mathbf{B}\left(\frac{1}{k}\sum_{i=1}^{k} P_{XY}^{(i)}\right)$. Then we have the following theorem.

Theorem 3.2 (EPR of Global Model Algorithm): With Assumption 2.1, the expected population risk as defined in (9) is

$$R_{\texttt{GM}} = kV_{\texttt{GM}} + \frac{|\mathcal{X}||\mathcal{Y}|}{n}\cdot O(1), \tag{17}$$

where

$$V_{\texttt{GM}} \triangleq \frac{1}{k}\sum_{i=1}^{k} \left\| \mathbf{B}\left(P_{XY}^{(i)}\right) - \mathbf{B}\left(\frac{1}{k}\sum_{j=1}^{k} P_{XY}^{(j)}\right) \right\|_{\mathrm{F}}^2. \tag{18}$$

This quantity $V_{\texttt{GM}}$ can be seen as the variance of different distributions, which does not decay to 0 with k. Overall, Theorem 3.2 states that the Global Model Algorithm provides a biased estimation and the variance term can be neglected under a large k.

3) Shared Representation Algorithm: For the Shared Representation Algorithm (Algorithm (iii)), the feature representation is learned with nk samples of all the devices, while each classifier is merely learned with its own n samples. We define

$$\boldsymbol{f}_{\texttt{SR}} \triangleq \underset{\boldsymbol{f}:\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{f})\boldsymbol{\Phi}(\boldsymbol{f})=\mathbf{I}_d}{\arg\max} \boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{f}) \left[\frac{1}{k}\sum_{i=1}^{k}\left(\boldsymbol{\Phi}_i\boldsymbol{\Psi}_i^{\mathrm{T}}\boldsymbol{\Psi}_i\boldsymbol{\Phi}_i^{\mathrm{T}}\right)\right] \boldsymbol{\Phi}(\boldsymbol{f}), \tag{19}$$

and the corresponding classifiers $\boldsymbol{g}_{\texttt{SR}i}$ as $\boldsymbol{\Psi}(\boldsymbol{g}_{\texttt{SR}i}) \triangleq \boldsymbol{\Psi}_i\boldsymbol{\Phi}_i^{\mathrm{T}}\boldsymbol{\Phi}(\boldsymbol{f}_{\texttt{SR}})$, for each $i = 1, \cdots, k$. Then, we can have the following characterization of the EPR (12).

[4] $\|\boldsymbol{\Phi}_i\|_{\mathrm{F}}^2$ corresponds to the summation of all the Hirschfeld–Gebelein–Rényi (HGR) correlation coefficients of $P_{XY}^{(i)}$, such that $\|\boldsymbol{\Phi}_i\|_{\mathrm{F}}^2 \le |\mathcal{Y}|$. [cf. Equation (30) [15]]

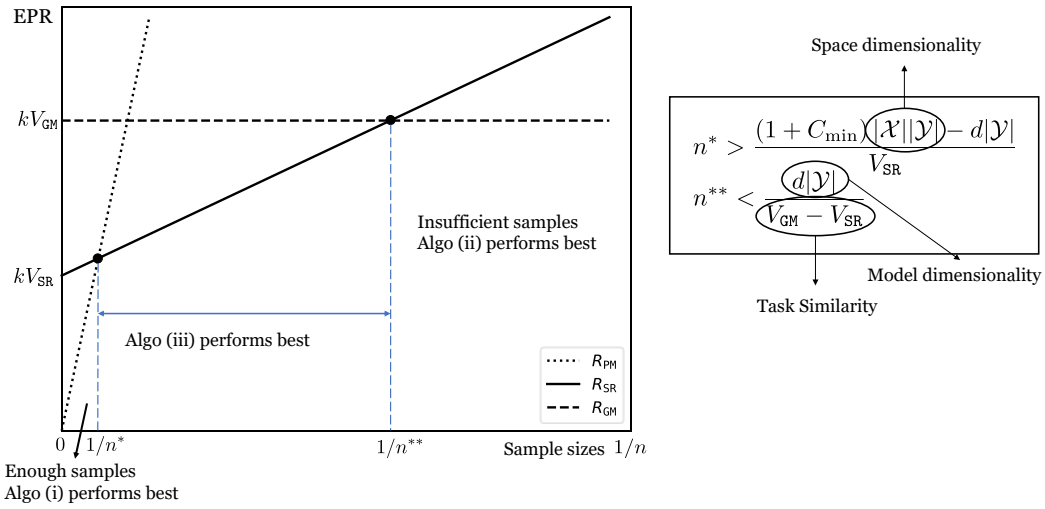

Fig. 2. A Summary of the EPRs as a function of $1/n$, which includes the Personalized Model Algorithm, the Global Model Algorithm, and the Shared Representation Algorithm. Sample sizes n^* and n^{**} represent the transition points that two algorithms reverse the orders in EPR comparisons.

Theorem 3.3 (EPR of Shared Representation Algorithm): With Assumption 2.1, the expected population risk as defined in (12) is

$$R_{\mathtt{SR}} = \underbrace{kV_{\mathtt{SR}}}_{\text{Bias}} + \underbrace{\frac{kd|\mathcal{Y}| - \sum_{i=1}^{k} \|\mathbf{\Psi}(\boldsymbol{g}_{\mathtt{SR}i})\|_{\mathrm{F}}^2}{n}}_{\text{Variance}} + \frac{|\mathcal{X}||\mathcal{Y}|}{n} \cdot O\left(1\right), \tag{20}$$

where

$$V_{\mathtt{SR}} \triangleq \frac{1}{k} \sum_{i=1}^{k} \left\| \mathbf{B} \left(P_{XY}^{(i)} \right) - \mathbf{\Psi}(\boldsymbol{g}_{\mathtt{SR}i}) \mathbf{\Phi}^{\mathrm{T}}(\boldsymbol{f}_{\mathtt{SR}}) \right\|_{\mathrm{F}}^2. \tag{21}$$

Theorem 3.3 states that the Shared Representation Algorithm has the EPR with both a bias term and a variance term. Specifically, the bias term is smaller than Algorithm (ii), and the variance term is smaller than Algorithm (i), which lead to the results in Section III-A.4.

4) Comparisons between the EPRs of Classical Algorithms: Based on the results above, we can have a summary of the mentioned EPRs, as shown in Figure 2. When neglecting the $O(1)$ terms, the expressions can be seen as functions of the sample size n. In view of the EPRs as linear functions of $1/n$, we set the abscissa as the reciprocal of the sample size.

Specifically, when the sample size n is large (as the left part of Figure 2), the data of the device allow training a well-performed model, where the Personalized Model Algorithm is preferable. When the sample size n is small (as the right part of Figure 2), the data of each device lead to a model with high error, where therefore the Global Model Algorithm is preferable. Moreover, under the following sample size range, the Shared Representation Algorithm can outperform the other two algorithms.

Corollary 3.4: When $n^{**} < n < n^*$, $R_{\mathtt{SR}} < R_{\mathtt{PM}}$ and $R_{\mathtt{SR}} < R_{\mathtt{GM}}$ are both satisfied, i.e., the Shared Representation Algorithm can perform best among the classical algorithms, where

$$\begin{aligned} n^* &= \frac{k|\mathcal{X}||\mathcal{Y}| - k - kd|\mathcal{Y}| + \sum_{i=1}^{k} \left(C_i + \|\mathbf{\Psi}(\boldsymbol{g}_{\mathtt{SR}i})\|_{\mathrm{F}}^2\right)}{kV_{\mathtt{SR}}} \\ &> \frac{(1+C_{\min})|\mathcal{X}||\mathcal{Y}| - d|\mathcal{Y}|}{V_{\mathtt{SR}}} \end{aligned} \tag{22}$$

and

$$n^{**} = \frac{kd|\mathcal{Y}| - \sum_{i=1}^{k} \|\mathbf{\Psi}(\boldsymbol{g}_{\mathtt{SR}i})\|_{\mathrm{F}}^2}{kV_{\mathtt{GM}} - kV_{\mathtt{SR}}} < \frac{d|\mathcal{Y}|}{V_{\mathtt{GM}} - V_{\mathtt{SR}}}. \tag{23}$$

Based on Corollary 3.4, we have the following interpretations of the transition points n^* and n^{**}. The transition point n^* is proportional to the *space dimensionality* $|\mathcal{X}||\mathcal{Y}|$ (considering the practical scene that $|\mathcal{X}|$ is much larger than d), and inversely proportional to the averaged distance between the underlying distributions and the learned models $V_{\mathtt{SR}}$. It indicates that a high-dimensional sample space and well-adapted model can lead to the preference of the Shared Representation Algorithm compared with the Personalized Model Algorithm.

On the other hand, the transition point n^{**} is proportional to $d|\mathcal{Y}|$, where it can be recognized as the *model dimensionality* of the classifiers with $d|\mathcal{Y}|$ training parameters. Meanwhile, the transition point is inversely proportional to $V_{\mathtt{GM}} - V_{\mathtt{SR}}$. When $\boldsymbol{f}_{\mathtt{SR}} = \boldsymbol{f}_{\mathtt{GM}}$, the numerator becomes $V_{\mathtt{GM}} - V_{\mathtt{SR}} = \mathrm{Var}\left(\{\boldsymbol{g}_{\mathtt{SR}i}\}\right) \triangleq \frac{1}{k}\sum_{i=1}^{k} \|\mathbf{\Psi}(\boldsymbol{g}_{\mathtt{SR}i}) - \mathbf{\Psi}(\frac{1}{k}\sum_{i=1}^{k} \boldsymbol{g}_{\mathtt{SR}i})\|_{\mathrm{F}}^2$, which can be seen as the variance of the learned classifiers and quantifies the *task similarity*. It explains that the low-dimensional model and the large task similarities lead to the preference of the Shared Representation Algorithm, in comparison with the Global Model Algorithm.

B. Classifiers Collaboration Algorithm

As introduced in Section II, the Classifiers Collaboration Algorithm (Algorithm (iv)) jointly employs the classifiers

learned by different devices. We prove in the following that its EPR can be declined through selecting proper combining coefficient α_i as defined in (13). As this algorithm is an operation based on Algorithm (iii), we do not provide the comparisons with Algorithm (i) or (ii) in this part.

First, for each device i, the EPR of the Shared Representation Algorithm as defined in (12) is

$$R_{\mathtt{SR}i} \triangleq R_i\left(\hat{\boldsymbol{f}}_{\mathtt{SR}}, \hat{\boldsymbol{g}}_{\mathtt{SR}i}\right) = \frac{d|\mathcal{Y}|}{n} - \frac{1}{n}\|\boldsymbol{\Psi}(\boldsymbol{g}_{\mathtt{SR}i})\|_{\mathrm{F}}^2 + \left\|\boldsymbol{\Psi}_i^{\mathrm{T}}\boldsymbol{\Phi}_i - \boldsymbol{\Psi}^{\mathrm{T}}(\boldsymbol{g}_{\mathtt{SR}i})\boldsymbol{\Phi}(\boldsymbol{f}_{\mathtt{SR}})\right\|_{\mathrm{F}}^2 + O\left(\frac{1}{nk}\right). \tag{24}$$

When α_i is a constant, we can compute the EPR (5) of each device as a function of α_i

$$\begin{aligned}\tilde{R}_i(\alpha_i) &\triangleq R_i\left(\hat{\boldsymbol{f}}_{\mathtt{SR}}, \alpha_i\hat{\boldsymbol{g}}_{\mathtt{SR}i} + (1-\alpha_i)\hat{\boldsymbol{g}}_{\mathtt{SR}}\right)\\ &= \alpha_i^2\left(\frac{d|\mathcal{Y}|}{n} - \frac{1}{n}\|\boldsymbol{\Psi}(\boldsymbol{g}_{\mathtt{SR}i})\|_{\mathrm{F}}^2\right) + (1-\alpha_i)^2\left\|\boldsymbol{\Psi}(\boldsymbol{g}_{\mathtt{SR}i}) - \boldsymbol{\Psi}(\boldsymbol{g}_{\mathtt{SR}})\right\|_{\mathrm{F}}^2\\ &\quad + \left\|\boldsymbol{\Psi}_i^{\mathrm{T}}\boldsymbol{\Phi}_i - \boldsymbol{\Psi}^{\mathrm{T}}(\boldsymbol{g}_{\mathtt{SR}i})\boldsymbol{\Phi}(\boldsymbol{f}_{\mathtt{SR}})\right\|_{\mathrm{F}}^2 + O\left(\frac{1}{nk}\right),\end{aligned} \tag{25}$$

where $\boldsymbol{g}_{\mathtt{SR}} = \frac{1}{k}\sum_{i=1}^k \boldsymbol{g}_{\mathtt{SR}i}$ and $\tilde{R}_i(1) = R_{\mathtt{SR}i}$. Thus, the optimal α_i^* is

$$\alpha_i^* = \arg\min_{\alpha_i} \tilde{R}_i(\alpha_i) = \frac{\|\boldsymbol{\Psi}(\boldsymbol{g}_{\mathtt{SR}i}) - \boldsymbol{\Psi}(\boldsymbol{g}_{\mathtt{SR}})\|_{\mathrm{F}}^2}{\|\boldsymbol{\Psi}(\boldsymbol{g}_{\mathtt{SR}i}) - \boldsymbol{\Psi}(\boldsymbol{g}_{\mathtt{SR}})\|_{\mathrm{F}}^2 + \frac{d|\mathcal{Y}|}{n} - \frac{1}{n}\|\boldsymbol{\Psi}(\boldsymbol{g}_{\mathtt{SR}i})\|_{\mathrm{F}}^2}. \tag{26}$$

Then we have the following characterization of the optimal coefficient α_i^*.

Proposition 3.5: The EPR as defined in (25) with respect to the parameter α_i^* as defined in (26) is

$$\tilde{R}_i(\alpha_i^*) = \frac{\left(\frac{d|\mathcal{Y}|}{n} - \frac{1}{n}\|\boldsymbol{\Psi}(\boldsymbol{g}_{\mathtt{SR}i})\|_{\mathrm{F}}^2\right)\|\boldsymbol{\Psi}(\boldsymbol{g}_{\mathtt{SR}i}) - \boldsymbol{\Psi}(\boldsymbol{g}_{\mathtt{SR}})\|_{\mathrm{F}}^2}{\frac{d|\mathcal{Y}|}{n} - \frac{1}{n}\|\boldsymbol{\Psi}(\boldsymbol{g}_{\mathtt{SR}i})\|_{\mathrm{F}}^2 + \|\boldsymbol{\Psi}(\boldsymbol{g}_{\mathtt{SR}i}) - \boldsymbol{\Psi}(\boldsymbol{g}_{\mathtt{SR}})\|_{\mathrm{F}}^2} + \left\|\boldsymbol{\Psi}_i^{\mathrm{T}}\boldsymbol{\Phi}_i - \boldsymbol{\Psi}^{\mathrm{T}}(\boldsymbol{g}_{\mathtt{SR}i})\boldsymbol{\Phi}(\boldsymbol{f}_{\mathtt{SR}})\right\|_{\mathrm{F}}^2 + O\left(\frac{1}{nk}\right) < R_{\mathtt{SR}i}. \tag{27}$$

Therefore, the Classifiers Collaboration Algorithm can provide a better performance than the Shared Representation Algorithm by selecting proper coefficient. In practice, the coefficient α_i^* also needs estimating. As the simplest way, we substitute the empirical classifiers for the true classifiers, i.e., we select

$$\alpha_i = \frac{\|\boldsymbol{\Psi}(\hat{\boldsymbol{g}}_{\mathtt{SR}i}) - \boldsymbol{\Psi}(\hat{\boldsymbol{g}}_{\mathtt{SR}})\|_{\mathrm{F}}^2}{\|\boldsymbol{\Psi}(\hat{\boldsymbol{g}}_{\mathtt{SR}i}) - \boldsymbol{\Psi}(\hat{\boldsymbol{g}}_{\mathtt{SR}})\|_{\mathrm{F}}^2 + \frac{d|\mathcal{Y}|}{n} - \frac{1}{n}\|\boldsymbol{\Psi}(\hat{\boldsymbol{g}}_{\mathtt{SR}i})\|_{\mathrm{F}}^2}. \tag{28}$$

However, due to the complicated expression of the undecided parameter α_i, the expectation of the population risk (14) is difficult to derive. Specifically, we consider the large sample size regime:

Assumption 3.6: The sample size n of each task satisfies $k^{\frac{1}{2}} > n \gg 1$.

As introduced in Section III-A.4, the right bound of the sample size interval that Algorithm (iii) performs best is proportional to the space dimensionality, where such assumption is often satisfied. Then we have the following characterization of $R_{\mathtt{CC}_i}$ as defined in (14).

Theorem 3.7 (EPR of Shared Representation Algorithm): With Assumption 2.1 and 3.6, the expected population risk as defined in (14) satisfies

$$R_{\mathtt{SR}i} - R_{\mathtt{CC}i} = (1-M_i)\frac{\left(d|\mathcal{Y}| - \|\boldsymbol{\Psi}(\boldsymbol{g}_{\mathtt{SR}i})\|_{\mathrm{F}}^2\right)^2}{n^2\|\boldsymbol{\Psi}(\boldsymbol{g}_{\mathtt{SR}i}) - \boldsymbol{\Psi}(\boldsymbol{g}_{\mathtt{SR}})\|_{\mathrm{F}}^2} + O\left(n^{-3}\right), \tag{29}$$

where M_i is defined in (48) of Appendix E.

The right-hand side of (29) decides whether the Classifiers Collaboration Algorithm is more preferable than the Shared Representation Algorithm. It is estimated in Appendix E that $M_i \lesssim \frac{4}{d-1}$. Thus, when $d > 5$, we have $R_{\mathtt{SR}i} > R_{\mathtt{CC}i}$, which means in the most practical scenes, the Classifiers Collaboration Algorithm can perform better than the Shared Representation Algorithm.

As a brief summary of this section, we present the EPRs of all the four algorithms and the assumptions of k and n in Table I.

IV. EPR Analyses: The Continuous case

Before this section, all the analyses are based on the sample space with a limited cardinality $|\mathcal{X}|$. However, for practical applications, such an assumption could lead to problems. Therefore, we extend the analyses above to the continuous case. The changes of our theoretical analyses mainly cover: (i) the representations of the feature extractors and classifiers are defined with parameterized models, (ii) the EPRs as functions of CDMs are transformed into integrals, and (iii) for analyses, the EPRs are computed under the large sample size regime.

For point (i), suppose that $\boldsymbol{w} = (\boldsymbol{w}_{\boldsymbol{f}}, \boldsymbol{w}_{\boldsymbol{g}})$ represents all the parameters to construct the feature extractor and classifier, where $\boldsymbol{w}_{\boldsymbol{f}}$ and $\boldsymbol{w}_{\boldsymbol{g}}$ represent the parameters contained in the feature extractor and classifier, respectively. In detail, the feature representation $\boldsymbol{f}(\boldsymbol{w}_{\boldsymbol{f}})$ expresses the function of data x with values $\boldsymbol{f}(x;\boldsymbol{w}_{\boldsymbol{f}})$, and similarly $\boldsymbol{g}(\boldsymbol{w}_{\boldsymbol{g}})$ expresses the function of label y with values $\boldsymbol{g}(y;\boldsymbol{w}_{\boldsymbol{g}})$. Specifically, the parameter vector $\boldsymbol{w}$ contains D parameters, i.e., $\boldsymbol{w} \in \mathbb{R}^D$.

For change (ii), the risk function with respect to CDMs can be computed in following way:

$$\left\|\mathbf{B}\left(\hat{P}_{XY}^{(i)}\right) - \boldsymbol{\Psi}(\boldsymbol{g}(\boldsymbol{w}_{\boldsymbol{g}}))\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{f}(\boldsymbol{w}_{\boldsymbol{f}}))\right\|_{\mathrm{F}}^2 = \int_{\mathcal{X}}\sum_{y\in\mathcal{Y}}\frac{\left(\hat{P}_{XY}^{(i)}(x,y) - Q_{XY}^{(\boldsymbol{w}_{\boldsymbol{f}},\boldsymbol{w}_{\boldsymbol{g}})}(x,y)\right)^2}{P_X(x)P_Y(y)}\mathrm{d}x, \tag{30}$$

where

$$Q_{XY}^{(\boldsymbol{w}_{\boldsymbol{f}},\boldsymbol{w}_{\boldsymbol{g}})}(x,y) \triangleq P_X(x)P_Y(y)\left(1 + \boldsymbol{f}^{\mathrm{T}}(x;\boldsymbol{w}_{\boldsymbol{f}})\boldsymbol{g}(y;\boldsymbol{w}_{\boldsymbol{g}})\right).$$

For change (iii), we employ Assumption 3.6. Under these changes, the difference of our main results lies in the EPR of the Personalized Model Algorithm. The parameters achieved

TABLE I

SUMMARY OF THE EPRS OF THE MENTIONED ALGORITHMS

EPR	Task number k	Sample size n	Expression
$R_{\texttt{PM}}$	Large	No request	$\frac{k\lvert\mathcal{X}\rvert\lvert\mathcal{Y}\rvert-k+\sum_{i=1}^{k}C_i}{n}$
$R_{\texttt{GM}}$	Large	No request	$kV_{\texttt{GM}}+\frac{\lvert\mathcal{X}\rvert\lvert\mathcal{Y}\rvert}{n}\cdot O\left(1\right)$
$R_{\texttt{SR}}$	Large	No request	$kV_{\texttt{SR}}+\frac{kd\lvert\mathcal{Y}\rvert-\sum_{i=1}^{k}\lVert\boldsymbol{\Psi}(\boldsymbol{g}_{\texttt{SR}i})\rVert_{\mathrm{F}}^2}{n}+\frac{\lvert\mathcal{X}\rvert\lvert\mathcal{Y}\rvert}{n}\cdot O\left(1\right)$
$R_{\texttt{CC}}$	Large	$k^{\frac{1}{2}}>n\gg 1$	$kV_{\texttt{SR}}+\frac{kd\lvert\mathcal{Y}\rvert-\sum_{i=1}^{k}\lVert\boldsymbol{\Psi}(\boldsymbol{g}_{\texttt{SR}i})\rVert_{\mathrm{F}}^2}{n}-\sum_{i=1}^{k}\left(1-M_i\right)\frac{\left(d\lvert\mathcal{Y}\rvert-\lVert\boldsymbol{\Psi}(\boldsymbol{g}_{\texttt{SR}i})\rVert_{\mathrm{F}}^2\right)^2}{n^2\lVert\boldsymbol{\Psi}(\boldsymbol{g}_{\texttt{SR}i})-\boldsymbol{\Psi}(\boldsymbol{g}_{\texttt{SR}})\rVert_{\mathrm{F}}^2}+O\left(\frac{k}{n^3}\right)$

TABLE II

SUMMARY OF THE EPRS OF THE MENTIONED ALGORITHMS UNDER THE CONTINUOUS DATA SPACE

EPR	Task number k	Sample size n	Expression
$R_{\texttt{PM}}$	Large	$k^{\frac{1}{2}}>n\gg 1$	$\sum_{i=1}^{k}\frac{D_i}{n}+O\left(\frac{k}{n^2}\right)$
$R_{\texttt{GM}}$	Large	$k^{\frac{1}{2}}>n\gg 1$	$kV_{\texttt{GM}}+O\left(\frac{1}{n}\right)$
$R_{\texttt{SR}}$	Large	$k^{\frac{1}{2}}>n\gg 1$	$kV_{\texttt{SR}}+\frac{kd\lvert\mathcal{Y}\rvert-\sum_{i=1}^{k}\lVert\boldsymbol{\Psi}(\boldsymbol{g}_{\texttt{SR}i})\rVert_{\mathrm{F}}^2}{n}+O\left(\frac{1}{n}\right)$
$R_{\texttt{CC}}$	Large	$k^{\frac{1}{2}}>n\gg 1$	$kV_{\texttt{SR}}+\frac{kd\lvert\mathcal{Y}\rvert-\sum_{i=1}^{k}\lVert\boldsymbol{\Psi}(\boldsymbol{g}_{\texttt{SR}i})\rVert_{\mathrm{F}}^2}{n}-\sum_{i=1}^{k}\left(1-M_i\right)\frac{\left(d\lvert\mathcal{Y}\rvert-\lVert\boldsymbol{\Psi}(\boldsymbol{g}_{\texttt{SR}i})\rVert_{\mathrm{F}}^2\right)^2}{n^2\lVert\boldsymbol{\Psi}(\boldsymbol{g}_{\texttt{SR}i})-\boldsymbol{\Psi}(\boldsymbol{g}_{\texttt{SR}})\rVert_{\mathrm{F}}^2}+O\left(\frac{k}{n^3}\right)$

can be defined with the feature and classifiers as in (6) such that

$$\hat{\boldsymbol{f}}_{\texttt{PM}i}=\boldsymbol{f}(\hat{\boldsymbol{w}}_{\boldsymbol{f}\mathrm{PM}i}),\ \text{and}\ \hat{\boldsymbol{g}}_{\texttt{PM}i}=\boldsymbol{g}(\hat{\boldsymbol{w}}_{\boldsymbol{g}\mathrm{PM}i}). \tag{31}$$

We define the risk

$$L_i(\boldsymbol{w})\triangleq\left\|\mathbf{B}(P_{XY}^{(i)})-\boldsymbol{\Phi}(\boldsymbol{g}(\boldsymbol{w}_{\boldsymbol{g}}))\boldsymbol{\Psi}^{\mathrm{T}}(\boldsymbol{f}(\boldsymbol{w}_{\boldsymbol{f}}))\right\|_{\mathrm{F}}^2.$$

Then, we have the following characterization of its EPR.

Theorem 4.1: With Assumption 2.1 and 3.6, suppose that there exist parameters $\boldsymbol{w}_{\mathrm{PM}i}=(\boldsymbol{w}_{\boldsymbol{f}\mathrm{PM}i},\boldsymbol{w}_{\boldsymbol{g}\mathrm{PM}i})$ such that $\boldsymbol{f}_i^{\mathrm{T}}(x)\boldsymbol{g}_i(y)=\boldsymbol{f}^{\mathrm{T}}(x;\boldsymbol{w}_{\boldsymbol{f}\mathrm{PM}i})\boldsymbol{g}(y;\boldsymbol{w}_{\boldsymbol{g}\mathrm{PM}i})$, for all $(x,y)\in\mathcal{X}\times\mathcal{Y}$ and let $D_i\triangleq\operatorname{rank}\left(\nabla_{\boldsymbol{w}}^2L_i(\boldsymbol{w})\big|_{\boldsymbol{w}=\boldsymbol{w}_{\mathrm{PM}i}}\right)$. The expected population risk as defined in (7) is

$$R_{\texttt{PM}}=\sum_{i=1}^{k}\frac{D_i}{n}+O\left(\frac{k}{n^2}\right)\leq\frac{kD}{n}. \tag{32}$$

The other ERPs of Algorithm (ii)-(iv) need slight modifications with respect to Assumption 3.6. For simplicity, we summarize all the results in Table II.

The transition point as defined in (22) can be estimated as $n^*\approx\frac{D}{V_{\texttt{SR}}}$, and the expression of n^{**} does not change. Similar interpretations can be given as in Section III-A.4, where n^* indicates that more parameters contained in the feature extractors can lead to the preference of the Shared Representation Algorithm compared with the Personalized Model Algorithm.

V. CONCLUSION

This paper introduces a mathematical framework for federate learning analyses under the large task number regime. We provide the closed-form EPR expressions of typical algorithms and characterize the roles of the task similarity and the model dimensionality, which explain the sample size range for employing the Shared Representation Algorithm. In addition, we propose an improved algorithm by collaborating the classifiers learned in different devices, whose effectiveness is verified by the EPR analyses.

APPENDIX

A. Proof of Theorem 3.1

According to the definition (7) and the fact $\boldsymbol{\Psi}(\hat{\boldsymbol{g}}_{\texttt{PM}i})\boldsymbol{\Phi}^{\mathrm{T}}(\hat{\boldsymbol{f}}_{\texttt{PM}i})=\mathbf{B}(\hat{P}_{XY}^{(i)})$, the EPR can be computed as

$$R_{\texttt{PM}}=\sum_{i=1}^{k}\mathbb{E}\left[\left\|\mathbf{B}\left(P_{XY}^{(i)}\right),\mathbf{B}\left(\hat{P}_{XY}^{(i)}\right)\right\|_{\mathrm{F}}^2\right]$$

$$
\begin{aligned}
&= \sum_{i=1}^{k} \mathbb{E}\left[\sum_{x\in\mathcal{X},y\in\mathcal{Y}} \frac{\left(P_{XY}^{(i)}(x,y) - \hat{P}_{XY}^{(i)}(x,y)\right)^2}{P_X(x)P_Y(y)}\right] \\
&= \sum_{i=1}^{k} \frac{1}{n} \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \frac{P_{XY}^{(i)}(x,y) - [P_{XY}^{(i)}(x,y)]^2}{P_X(x)P_Y(y)} \\
&= \sum_{i=1}^{k} \frac{1}{n} \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \big[(1 + \boldsymbol{f}_i^{\mathrm{T}}(x)\boldsymbol{g}_i(y)) \\
&\qquad - P_X(x)P_Y(y)(1 + \boldsymbol{f}_i^{\mathrm{T}}(x)\boldsymbol{g}_i(y))^2\big]. \qquad (33)
\end{aligned}
$$

With $\mathbb{E}_{P_X}[\boldsymbol{f}_i(X)] = \mathbb{E}_{P_Y}[\boldsymbol{g}_i(X)] = \mathbf{0}$ and $\boldsymbol{\Phi}_i^{\mathrm{T}}\boldsymbol{\Phi}_i = \mathbf{I}_d$, we can derive the theorem.

B. Proof of Theorem 3.2

According to the definition (9) and the fact $\boldsymbol{\Psi}(\hat{\boldsymbol{g}}_{\mathtt{GM}})\boldsymbol{\Phi}^{\mathrm{T}}(\hat{\boldsymbol{f}}_{\mathtt{GM}}) = \mathbf{B}\left(\frac{1}{k}\sum_{i=1}^{k}\hat{P}_{XY}^{(i)}\right)$, the EPR can be computed as

$$
\begin{aligned}
R_{\mathtt{GM}} &= \sum_{i=1}^{k} \mathbb{E}\left[\left\|\mathbf{B}\left(P_{XY}^{(i)}\right), \mathbf{B}\left(\frac{1}{k}\sum_{i=1}^{k}\hat{P}_{XY}^{(i)}\right)\right\|_{\mathrm{F}}^2\right] \\
&= \sum_{i=1}^{k} \mathbb{E}\left[\sum_{x\in\mathcal{X},y\in\mathcal{Y}} \frac{\left(P_{XY}^{(i)}(x,y) - \frac{1}{k}\sum_{j=1}^{k}\hat{P}_{XY}^{(j)}(x,y)\right)^2}{P_X(x)P_Y(y)}\right] \\
&= \sum_{i=1}^{k} \mathbb{E}\left[\sum_{x\in\mathcal{X},y\in\mathcal{Y}} \frac{\left(P_{XY}^{(i)}(x,y) - \frac{1}{k}\sum_{j=1}^{k}P_{XY}^{(j)}(x,y)\right)^2}{P_X(x)P_Y(y)}\right] \\
&\quad + \sum_{i=1}^{k} \mathbb{E}\left[\sum_{x\in\mathcal{X},y\in\mathcal{Y}} \frac{\left(\sum_{j=1}^{k}\hat{P}_{XY}^{(j)}(x,y) - \sum_{j=1}^{k}P_{XY}^{(j)}(x,y)\right)^2}{k^2P_X(x)P_Y(y)}\right], \qquad (34)
\end{aligned}
$$

where the last equation comes from the fact that $\mathbb{E}[\hat{P}_{XY}^{(j)}] = P_{XY}^{(j)}$. With the result in Section A, we have

$$
\begin{aligned}
&\sum_{i=1}^{k} \mathbb{E}\left[\sum_{x\in\mathcal{X},y\in\mathcal{Y}} \frac{\left(\sum_{j=1}^{k}\hat{P}_{XY}^{(j)}(x,y) - \sum_{j=1}^{k}P_{XY}^{(j)}(x,y)\right)^2}{k^2P_X(x)P_Y(y)}\right] \\
&= \frac{|\mathcal{X}||\mathcal{Y}| \cdot O(1)}{n}.
\end{aligned}
$$

Then, Theorem 3.2 is proved.

C. Proof of Theorem 3.3

Note that

$$
\hat{\boldsymbol{g}}_{\mathtt{SR}i} = \arg\min_{\boldsymbol{g}} \left\|\mathbf{B}\left(\hat{P}_{XY}^{(i)}\right) - \boldsymbol{\Psi}(\boldsymbol{g})\boldsymbol{\Phi}^{\mathrm{T}}(\hat{\boldsymbol{f}}_{\mathtt{SR}})\right\|_{\mathrm{F}}^2. \qquad (35)
$$

We can compute by the derivatives that

$$
\boldsymbol{\Psi}(\hat{\boldsymbol{g}}_{\mathtt{SR}i}) = \mathbf{B}\left(\hat{P}_{XY}^{(i)}\right)\boldsymbol{\Phi}(\hat{\boldsymbol{f}}_{\mathtt{SR}}). \qquad (36)
$$

With

$$
\hat{\boldsymbol{f}}_{\mathtt{SR}} = \arg\min_{\boldsymbol{f}:\boldsymbol{\Psi}^{\mathrm{T}}(\boldsymbol{f})\boldsymbol{\Psi}(\boldsymbol{f})=\mathbf{I}_d} \sum_{i=1}^{k} \left\|\mathbf{B}\left(\hat{P}_{XY}^{(i)}\right) - \boldsymbol{\Psi}(\hat{\boldsymbol{g}}_{\mathtt{SR}i})\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{f})\right\|_{\mathrm{F}}^2, \qquad (37)
$$

we have

$$
\begin{aligned}
\hat{\boldsymbol{f}}_{\mathtt{SR}} = &\arg\max_{\boldsymbol{f}:\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{f})\boldsymbol{\Phi}(\boldsymbol{f})=\mathbf{I}_d} \\
&\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{f})\left[\frac{1}{k}\sum_{i=1}^{k}\mathbf{B}^{\mathrm{T}}\left(\hat{P}_{XY}^{(i)}\right)\mathbf{B}\left(\hat{P}_{XY}^{(i)}\right)\right]\boldsymbol{\Phi}(\boldsymbol{f}).
\end{aligned}
$$

Note that (19) defines

$$
\begin{aligned}
\boldsymbol{f}_{\mathtt{SR}} \triangleq &\arg\max_{\boldsymbol{f}:\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{f})\boldsymbol{\Phi}(\boldsymbol{f})=\mathbf{I}_d} \\
&\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{f})\left[\frac{1}{k}\sum_{i=1}^{k}\mathbf{B}^{\mathrm{T}}\left(P_{XY}^{(i)}\right)\mathbf{B}\left(P_{XY}^{(i)}\right)\right]\boldsymbol{\Phi}(\boldsymbol{f}).
\end{aligned}
$$

For simplicity, let $\mathbf{A} \triangleq \frac{1}{k}\sum_{i=1}^{k}\mathbf{B}^{\mathrm{T}}\left(P_{XY}^{(i)}\right)\mathbf{B}\left(P_{XY}^{(i)}\right)$ and $\hat{\mathbf{A}} \triangleq \frac{1}{k}\sum_{i=1}^{k}\mathbf{B}^{\mathrm{T}}\left(\hat{P}_{XY}^{(i)}\right)\mathbf{B}\left(\hat{P}_{XY}^{(i)}\right)$.

(i) When $\mathbf{A} = \hat{\mathbf{A}}$, $\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{f}_{\mathtt{SR}}) = \boldsymbol{\Phi}^{\mathrm{T}}(\hat{\boldsymbol{f}}_{\mathtt{SR}})$.

(ii) When $\mathbf{A} \neq \hat{\mathbf{A}}$, according to Davis-Kahan's Theorem [21], we have

$$
\begin{aligned}
\frac{1}{2}\left\|\boldsymbol{\Phi}(\boldsymbol{f}_{\mathtt{SR}}) - \boldsymbol{\Phi}(\hat{\boldsymbol{f}}_{\mathtt{SR}})\right\|_{\mathrm{F}}^2 &\leq \left\|\sin\Theta\left(\boldsymbol{\Phi}(\boldsymbol{f}_{\mathtt{SR}}), \boldsymbol{\Phi}(\hat{\boldsymbol{f}}_{\mathtt{SR}})\right)\right\|_{\mathrm{F}}^2 \\
&\leq \frac{\left\|\mathbf{A} - \hat{\mathbf{A}}\right\|_{\mathrm{F}}^2}{\delta}, \qquad (38)
\end{aligned}
$$

where the eigenvalues of $\mathbf{A}$ and $\hat{\mathbf{A}}$ are $\lambda_1 \geq \cdots \geq \lambda_{|\mathcal{X}|}$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_{|\mathcal{X}|}$, and

$$
\delta \triangleq \inf_{\hat{\mathbf{A}}} \inf_{1\leq i\leq d, d\leq j\leq|\mathcal{X}|} \left|\lambda_i - \hat{\lambda}_j\right|. \qquad (39)
$$

It is easy to verify that

$$
\mathbb{E}\left[\left\|\mathbf{A} - \hat{\mathbf{A}}\right\|_{\mathrm{F}}^2\right] = \frac{|\mathcal{X}||\mathcal{Y}|}{n} \cdot O\left(k^{-1}\right). \qquad (40)
$$

Considering the cases (i) and (ii), we have

$$
\mathbb{E}\left[\left\|\boldsymbol{\Phi}(\boldsymbol{f}_{\mathtt{SR}}) - \boldsymbol{\Phi}(\hat{\boldsymbol{f}}_{\mathtt{SR}})\right\|_{\mathrm{F}}^2\right] = \frac{|\mathcal{X}||\mathcal{Y}|}{n} \cdot O\left(k^{-1}\right). \qquad (41)
$$

Now, we can compute the EPR (12):

$$
\begin{aligned}
&R_{\mathtt{SR}} \\
&= \sum_{i=1}^{k} \mathbb{E}\left[\left\|\mathbf{B}\left(P_{XY}^{(i)}\right) - \mathbf{B}\left(\hat{P}_{XY}^{(i)}\right)\mathbf{\Phi}(\hat{\boldsymbol{f}}_{\mathtt{SR}})\mathbf{\Phi}^{\mathrm{T}}(\hat{\boldsymbol{f}}_{\mathtt{SR}})\right\|_{\mathrm{F}}^{2}\right] \\
&= \sum_{i=1}^{k} \mathbb{E}\left[\left\|\mathbf{B}\left(P_{XY}^{(i)}\right) - \mathbf{B}\left(\hat{P}_{XY}^{(i)}\right)\mathbf{\Phi}(\boldsymbol{f}_{\mathtt{SR}})\mathbf{\Phi}^{\mathrm{T}}(\boldsymbol{f}_{\mathtt{SR}})\right\|_{\mathrm{F}}^{2}\right] \\
&\quad + \frac{|\mathcal{X}||\mathcal{Y}|}{n}\cdot O\left(1\right) \\
&= \sum_{i=1}^{k} \mathbb{E}\left[\left\|\mathbf{B}\left(P_{XY}^{(i)}\right) - \mathbf{B}\left(P_{XY}^{(i)}\right)\mathbf{\Phi}(\boldsymbol{f}_{\mathtt{SR}})\mathbf{\Phi}^{\mathrm{T}}(\boldsymbol{f}_{\mathtt{SR}})\right\|_{\mathrm{F}}^{2}\right] \\
&\quad + \sum_{i=1}^{k} \mathbb{E}\left[\left\|\mathbf{B}\left(P_{XY}^{(i)} - \hat{P}_{XY}^{(i)}\right)\mathbf{\Phi}(\boldsymbol{f}_{\mathtt{SR}})\mathbf{\Phi}^{\mathrm{T}}(\boldsymbol{f}_{\mathtt{SR}})\right\|_{\mathrm{F}}^{2}\right] \\
&\quad + \frac{|\mathcal{X}||\mathcal{Y}|}{n}\cdot O\left(1\right) \\
&= kV_{\mathtt{SR}} + \sum_{i=1}^{k}\frac{1}{n}\sum_{y=1}^{|\mathcal{Y}|}\mathrm{tr}\left(\mathbb{E}_{P_{X|Y=y}^{(i)}}[\boldsymbol{f}_{\mathtt{SR}}(X)\boldsymbol{f}_{\mathtt{SR}}^{\mathrm{T}}(X)]\right) \\
&\quad - \sum_{i=1}^{k}\frac{1}{n}\sum_{y=1}^{|\mathcal{Y}|}P_Y(y)\left\|\mathbb{E}_{P_{X|Y=y}^{(i)}}[\boldsymbol{f}_{\mathtt{SR}}(X)]\right\|^{2} \\
&\quad + \frac{|\mathcal{X}||\mathcal{Y}|}{n}\cdot O\left(1\right) \\
&= kV_{\mathtt{SR}} + \frac{kd|\mathcal{Y}| - \sum_{i=1}^{k}\|\mathbf{\Psi}(\boldsymbol{g}_{\mathtt{SR}i})\|_{\mathrm{F}}^{2}}{n} + \frac{|\mathcal{X}||\mathcal{Y}|}{n}\cdot O\left(1\right).
\end{aligned}
\tag{42}
$$

Then, Theorem 3.3 is proved.

D. Local Approximation of the multinomial distribution

According to the multinomial distribution,

$$
\mathbb{P}\left(\hat{P}_{XY}^{(i)}; P_{XY}^{(i)}\right) \doteq \exp\left(-nD\left(\hat{P}_{XY}^{(i)}\|P_{XY}^{(i)}\right)\right). \tag{43}
$$

Moreover, we introduce the following lemma:

Lemma 5.1: Suppose that $|P(z) - Q(z)| < \epsilon$ for each $z \in \mathcal{Z}$, where $\epsilon/|\mathcal{Z}| \ll 1$, K-L divergence between $P(z)$ and $Q(z)$ can have the following approximation

$$
D\left(P\|Q\right) = \frac{1}{2}\chi^2(P,Q) + O(\epsilon^3). \tag{44}
$$

With (43) and Lemma 5.1, we let

$$
\mathbb{P}\left(\hat{P}_{XY}^{(i)}; P_{XY}^{(i)}\right) \propto \exp\left(-\frac{n}{2}\chi^2\left(P_{XY}^{(i)}, \hat{P}_{XY}^{(i)}\right)\right). \tag{45}
$$

E. Proof of Theorem 3.7

First, we prove the EPR with constant α_i as define in (25).

$$
\begin{aligned}
&\tilde{R}_i(\alpha_i) \triangleq R_i\left(\hat{\boldsymbol{f}}_{\mathtt{SR}}, \alpha_i\hat{\boldsymbol{g}}_{\mathtt{SR}i} + (1-\alpha_i)\hat{\boldsymbol{g}}_{\mathtt{SR}}\right) \\
&= \mathbb{E}\left[\left\|\mathbf{B}\left(P_{XY}^{(i)}\right) - \left(\alpha_i\mathbf{B}\left(\hat{P}_{XY}^{(i)}\right)\mathbf{\Phi}(\hat{\boldsymbol{f}}_{\mathtt{SR}})\right.\right.\right. \\
&\quad \left.\left.\left. + (1-\alpha_i)\mathbf{B}\Big(\frac{1}{k}\sum_{j=1}^{k}\hat{P}_{XY}^{(j)}\Big)\mathbf{\Phi}(\hat{\boldsymbol{f}}_{\mathtt{SR}})\Big)\mathbf{\Phi}^{\mathrm{T}}(\hat{\boldsymbol{f}}_{\mathtt{SR}})\right\|_{\mathrm{F}}^{2}\right] \\
&= \mathbb{E}\left[\left\|\mathbf{B}\left(P_{XY}^{(i)}\right) - \left(\alpha_i\mathbf{B}\left(\hat{P}_{XY}^{(i)}\right)\mathbf{\Phi}(\hat{\boldsymbol{f}}_{\mathtt{SR}})\right.\right.\right. \\
&+ \left.\left.\left. (1-\alpha_i)\mathbf{B}\Big(\frac{1}{k}\sum_{j=1}^{k}P_{XY}^{(j)}\Big)\mathbf{\Phi}(\boldsymbol{f}_{\mathtt{SR}})\Big)\mathbf{\Phi}^{\mathrm{T}}(\boldsymbol{f}_{\mathtt{SR}})\right\|_{\mathrm{F}}^{2}\right] + O\left(\frac{1}{nk}\right) \\
&= \alpha_i^2\,\mathbb{E}\left[\left\|\mathbf{B}\left(P_{XY}^{(i)} - \hat{P}_{XY}^{(i)}\right)\mathbf{\Phi}(\boldsymbol{f}_{\mathtt{SR}})\mathbf{\Phi}^{\mathrm{T}}(\boldsymbol{f}_{\mathtt{SR}})\right\|_{\mathrm{F}}^{2}\right] \\
&\quad + (1-\alpha_i)^2\left\|\mathbf{B}\left(P_{XY}^{(i)} - \frac{1}{k}\sum_{j=1}^{k}P_{XY}^{(j)}\right)\mathbf{\Phi}(\boldsymbol{f}_{\mathtt{SR}})\mathbf{\Phi}^{\mathrm{T}}(\boldsymbol{f}_{\mathtt{SR}})\right\|_{\mathrm{F}}^{2} \\
&\quad + \left\|\mathbf{B}\left(P_{XY}^{(i)}\right) - \mathbf{B}\left(P_{XY}^{(i)}\right)\mathbf{\Phi}(\boldsymbol{f}_{\mathtt{SR}})\mathbf{\Phi}^{\mathrm{T}}(\boldsymbol{f}_{\mathtt{SR}})\right\|_{\mathrm{F}}^{2} + O\left(\frac{1}{nk}\right) \\
&= \alpha_i^2\left(\frac{d|\mathcal{Y}|}{n} - \frac{1}{n}\|\mathbf{\Psi}(\boldsymbol{g}_{\mathtt{SR}i})\|_{\mathrm{F}}^{2}\right) \\
&\quad + (1-\alpha_i)^2\left\|\mathbf{\Psi}(\boldsymbol{g}_{\mathtt{SR}i}) - \mathbf{\Psi}(\boldsymbol{g}_{\mathtt{SR}})\right\|_{\mathrm{F}}^{2} \\
&\quad + \left\|\mathbf{\Psi}_i^{\mathrm{T}}\mathbf{\Phi}_i - \mathbf{\Psi}^{\mathrm{T}}(\boldsymbol{g}_{\mathtt{SR}i})\mathbf{\Phi}(\boldsymbol{f}_{\mathtt{SR}})\right\|_{\mathrm{F}}^{2} + O\left(\frac{1}{nk}\right).
\end{aligned}
\tag{46}
$$

Accordingly, we can derive α_i^*. With Assumption 3.6 and $\Delta\boldsymbol{g}_{\mathtt{SR}i} \triangleq \hat{\boldsymbol{g}}_{\mathtt{SR}i} - \boldsymbol{g}_{\mathtt{SR}i}$,

$$
\begin{aligned}
\alpha_i &= \frac{\|\mathbf{\Psi}(\hat{\boldsymbol{g}}_{\mathtt{SR}i}) - \mathbf{\Psi}(\hat{\boldsymbol{g}}_{\mathtt{SR}})\|_{\mathrm{F}}^{2}}{\|\mathbf{\Psi}(\hat{\boldsymbol{g}}_{\mathtt{SR}i}) - \mathbf{\Psi}(\hat{\boldsymbol{g}}_{\mathtt{SR}})\|_{\mathrm{F}}^{2} + \frac{d|\mathcal{Y}|}{n} - \frac{1}{n}\|\mathbf{\Psi}(\hat{\boldsymbol{g}}_{\mathtt{SR}i})\|_{\mathrm{F}}^{2}} \\
&= \alpha_i^* + O(\|\mathbf{\Psi}(\Delta\boldsymbol{g}_{\mathtt{SR}i}\|_{\mathrm{F}}^{2}) \\
&\quad + 2\frac{\langle\theta_i\mathbf{\Psi}(\boldsymbol{g}_{\mathtt{SR}i} - \boldsymbol{g}_{\mathtt{SR}}) + \theta\mathbf{\Psi}(\boldsymbol{g}_{\mathtt{SR}i}), \mathbf{\Psi}(\Delta\boldsymbol{g}_{\mathtt{SR}i})\rangle_{\mathrm{F}}}{n\theta^2},
\end{aligned}
$$

where

$$
\theta = \|\mathbf{\Psi}(\boldsymbol{g}_{\mathtt{SR}i} - \boldsymbol{g}_{\mathtt{SR}})\|_{\mathrm{F}}^{2},
$$

and

$$
\theta_i = d|y| - \|\mathbf{\Psi}(\boldsymbol{g}_{\mathtt{SR}i})\|_{\mathrm{F}}^{2}
$$

Then, we can have Theorem 3.7 as follows

$$
\begin{aligned}
&R_i\left(\hat{\boldsymbol{f}}_{\mathtt{SR}}, \hat{\boldsymbol{g}}_{\mathtt{CC}i}\right) \\
&= \mathbb{E}\left[\left\|\mathbf{B}\left(P_{XY}^{(i)}\right) - \left(\alpha_i\mathbf{B}\left(\hat{P}_{XY}^{(i)}\right)\mathbf{\Phi}(\hat{\boldsymbol{f}}_{\mathtt{SR}})\right.\right.\right. \\
&\quad \left.\left.\left. + (1-\alpha_i)\mathbf{B}\Big(\frac{1}{k}\sum_{j=1}^{k}\hat{P}_{XY}^{(j)}\Big)\mathbf{\Phi}(\hat{\boldsymbol{f}}_{\mathtt{SR}})\Big)\mathbf{\Phi}^{\mathrm{T}}(\hat{\boldsymbol{f}}_{\mathtt{SR}})\right\|_{\mathrm{F}}^{2}\right] \\
&= \tilde{R}_i(\alpha_i^*) + O\left(\frac{1}{n^3}\right) \\
&\quad + 4\,\mathbb{E}\left[\alpha_i^*\frac{\langle\theta_i\mathbf{\Psi}(\boldsymbol{g}_{\mathtt{SR}i} - \boldsymbol{g}_{\mathtt{SR}}) + \theta\mathbf{\Psi}(\boldsymbol{g}_{\mathtt{SR}i}), \mathbf{\Psi}(\Delta\boldsymbol{g}_{\mathtt{SR}i})\rangle_{\mathrm{F}}}{n\theta^2}\right.
\end{aligned}
$$

$$\left\langle \boldsymbol{\Psi}(\boldsymbol{g}_{\mathtt{SR}i} - \boldsymbol{g}_{\mathtt{SR}}), \boldsymbol{\Psi}(\Delta \boldsymbol{g}_{\mathtt{SR}i}) \right\rangle_{\mathrm{F}} \Bigg]$$
$$= \tilde{R}_i(\alpha_i^*) + M_i \frac{\theta_i^2}{n^2\theta} + O\left(\frac{1}{n^3}\right), \tag{47}$$

where

$$M_i = \frac{4}{\theta\theta_i}\left(\sum_{x,y} P_{XY}^{(i)}(x,y)\boldsymbol{f}_{\mathtt{SR}}^{\mathrm{T}}(x)(\boldsymbol{h}_1(y)+\boldsymbol{h}_2(y))\boldsymbol{f}_{\mathtt{SR}}^{\mathrm{T}}(x)\boldsymbol{h}_1(y) - \sum_y P_Y(y)\boldsymbol{h}_1^{\mathrm{T}}(y)\boldsymbol{g}_{\mathtt{SR}i}(y) \cdot \sum_y P_Y(y)\boldsymbol{h}_2^{\mathrm{T}}(y)\boldsymbol{g}_{\mathtt{SR}i}(y)\right) \tag{48}$$

where $\boldsymbol{h}_1 \triangleq \boldsymbol{g}_{\mathtt{SR}i} - \boldsymbol{g}_{\mathtt{SR}}$ and $\boldsymbol{h}_2 \triangleq \frac{\theta}{\theta_i}\boldsymbol{g}_{\mathtt{SR}i}$.

Consider that $\boldsymbol{g}_{\mathtt{SR}i} - \boldsymbol{g}_{\mathtt{SR}}$ and $\boldsymbol{g}_i$ can have a positive inner product or negative inner product. We apply that in average $\mathbb{E}[\boldsymbol{h}_1^{\mathrm{T}}(Y)\boldsymbol{g}_{\mathtt{SR}i}(Y)] \approx 0$. Note that $\mathbb{E}[\boldsymbol{f}_{\mathtt{SR}}(X)\boldsymbol{f}_{\mathtt{SR}}^{\mathrm{T}}(X)] = \boldsymbol{I}$. Then we can estimate M_i as

$$\begin{aligned} M_i &\approx \frac{4}{\theta\theta_i}\sum_{x,y} P_{XY}^{(i)}(x,y)\boldsymbol{f}_{\mathtt{SR}}^{\mathrm{T}}(x)(\boldsymbol{h}_1(y)+\boldsymbol{h}_2(y))\boldsymbol{f}_{\mathtt{SR}}^{\mathrm{T}}(x)\boldsymbol{h}_1(y) \\ &\le \frac{4}{\theta\theta_i}|\mathcal{Y}|\sum_y P_Y(y)(\boldsymbol{h}_1(y)+\boldsymbol{h}_2(y))^{\mathrm{T}}\boldsymbol{h}_1(y) \\ &\approx \frac{4|\mathcal{Y}|}{\theta_i} \le \frac{4|\mathcal{Y}|}{d|\mathcal{Y}| - |\mathcal{Y}|} = \frac{4}{d-1} \end{aligned} \tag{49}$$

Note that

$$\tilde{R}_i(\alpha_i^*) - \tilde{R}_i(\alpha_i^*) = \frac{\theta_i^2}{n^2\theta} + O\left(\frac{1}{n^3}\right). \tag{50}$$

We can easily derive Theorem 3.7.

F. Proof of Theorem 4.1

Note that the risk function is $L_i(\boldsymbol{w}) = \left\|\mathbf{B}(P_{XY}^{(i)}) - \boldsymbol{\Phi}(\boldsymbol{g}(\boldsymbol{w}_{\boldsymbol{g}}))\boldsymbol{\Psi}^{\mathrm{T}}(\boldsymbol{f}(\boldsymbol{w}_{\boldsymbol{f}}))\right\|_{\mathrm{F}}^2$. We also define the empirical risk

$$l_i(\boldsymbol{p}, \boldsymbol{w}) = \left\|\mathbf{B}(\boldsymbol{p}) - \boldsymbol{\Phi}(\boldsymbol{g}(\boldsymbol{w}_{\boldsymbol{g}}))\boldsymbol{\Psi}^{\mathrm{T}}(\boldsymbol{f}(\boldsymbol{w}_{\boldsymbol{f}}))\right\|_{\mathrm{F}}^2, \tag{51}$$

and $L_i(\boldsymbol{w}) = l_i(P_{XY}^{(i)}, \boldsymbol{w})$. With $\boldsymbol{w}_{\mathrm{PM}i} = (\boldsymbol{w}_{\boldsymbol{f}\mathrm{PM}i}, \boldsymbol{w}_{\boldsymbol{g}\mathrm{PM}i})$ and $\hat{\boldsymbol{w}}_{\mathrm{PM}i} = (\hat{\boldsymbol{w}}_{\boldsymbol{f}\mathrm{PM}i}, \hat{\boldsymbol{w}}_{\boldsymbol{g}\mathrm{PM}i})$

$$\nabla_{\boldsymbol{w}} l_i(P_{XY}^{(i)}, \boldsymbol{w}_{\mathrm{PM}i}) = \nabla_{\boldsymbol{w}} l_i(\hat{P}_{XY}^{(i)}, \hat{\boldsymbol{w}}_{\mathrm{PM}i}) = 0. \tag{52}$$

By the Taylor expansions of $\nabla_{\boldsymbol{w}} l_i(\boldsymbol{p}, \boldsymbol{w})$, let $\Delta \boldsymbol{p}_i \triangleq \hat{P}_{XY}^{(i)} - P_{XY}^{(i)} \in \mathbb{R}^{|\mathcal{X}||\mathcal{Y}|}$, $\Delta \boldsymbol{w}_i \triangleq \hat{\boldsymbol{w}}_{\mathrm{PM}i} - \boldsymbol{w}_{\mathrm{PM}i}$, and then we have

$$\Delta \boldsymbol{w}_i = \left[\nabla_{\boldsymbol{w}}^2 l_i(P_{XY}^{(i)}, \boldsymbol{w}_{\mathrm{PM}i})\right]^{-1} \cdot \left[\nabla_{\boldsymbol{w},\boldsymbol{p}}^2 l_i(P_{XY}^{(i)}, \boldsymbol{w}_{\mathrm{PM}i})\right] \Delta \boldsymbol{p}_i + O\left(\|\Delta \boldsymbol{p}_i\|^2\right). \tag{53}$$

Then, the EPR is

$$\begin{aligned} R_i\left(\hat{\boldsymbol{f}}_{\mathrm{PM}i}, \hat{\boldsymbol{g}}_{\mathrm{PM}i}\right) &= \mathbb{E}[L_i(\hat{\boldsymbol{w}}_{\mathrm{PM}i})] \\ &= \frac{1}{2}\mathbb{E}\left[\Delta \boldsymbol{p}_i^{\mathrm{T}}\left[\nabla_{\boldsymbol{w},\boldsymbol{p}}^2 l_i(P_{XY}^{(i)}, \boldsymbol{w}_{\mathrm{PM}i})\right]^{\mathrm{T}} \left[\nabla_{\boldsymbol{w}}^2 l_i(P_{XY}^{(i)}, \boldsymbol{w}_{\mathrm{PM}i})\right]^{-1} \left[\nabla_{\boldsymbol{w},\boldsymbol{p}}^2 l_i(P_{XY}^{(i)}, \boldsymbol{w}_{\mathrm{PM}i})\right] \Delta \boldsymbol{p}_i\right] \\ &\quad + \mathbb{E}\left[O\left(\|\Delta \boldsymbol{p}_i\|^3\right)\right] \\ &= \frac{1}{2}\left[\nabla_{\boldsymbol{w},\boldsymbol{p}}^2 l_i(P_{XY}^{(i)}, \boldsymbol{w}_{\mathrm{PM}i})\right]\mathbb{E}\left[\Delta \boldsymbol{p}_i \Delta \boldsymbol{p}_i^{\mathrm{T}}\right] \left[\nabla_{\boldsymbol{w},\boldsymbol{p}}^2 l_i(P_{XY}^{(i)}, \boldsymbol{w}_{\mathrm{PM}i})\right]^{\mathrm{T}} \left[\nabla_{\boldsymbol{w}}^2 l_i(P_{XY}^{(i)}, \boldsymbol{w}_{\mathrm{PM}i})\right]^{-1} \\ &\quad + O\left(\frac{1}{n^2}\right) \\ &= \frac{1}{2n}\operatorname{tr}\left(\left[\nabla_{\boldsymbol{w},\boldsymbol{p}}^2 l_i(P_{XY}^{(i)}, \boldsymbol{w}_{\mathrm{PM}i})\right]\operatorname{diag}\left\{P_{XY}^{(i)}\right\} \left[\nabla_{\boldsymbol{w},\boldsymbol{p}}^2 l_i(P_{XY}^{(i)}, \boldsymbol{w}_{\mathrm{PM}i})\right]^{\mathrm{T}} \left[\nabla_{\boldsymbol{w}}^2 l_i(P_{XY}^{(i)}, \boldsymbol{w}_{\mathrm{PM}i})\right]^{-1}\right) \\ &\quad + O\left(\frac{1}{n^2}\right). \end{aligned} \tag{54}$$

When $l_i(P_{XY}^{(i)}, \boldsymbol{w}_{\mathrm{PM}i}) = 0$, let $\boldsymbol{q}(x, y; \boldsymbol{w}) = \boldsymbol{f}^{\mathrm{T}}(x; \boldsymbol{w}_{\boldsymbol{f}})\boldsymbol{g}(y; \boldsymbol{w}_{\boldsymbol{g}\mathrm{PM}i})$, and then we have

$$\begin{aligned} &\nabla_{\boldsymbol{w}}^2 l_i(P_{XY}^{(i)}, \boldsymbol{w}_{\mathrm{PM}i}) \\ &= 2\,\mathbb{E}_{P_{XY}^{(i)}}[\nabla_{\boldsymbol{w}}\boldsymbol{q}(X, Y; \boldsymbol{w}_{\mathrm{PM}i})\nabla_{\boldsymbol{w}}\boldsymbol{q}(X, Y; \boldsymbol{w}_{\mathrm{PM}i})^{\mathrm{T}}], \end{aligned} \tag{55}$$

$$\begin{aligned} &\left[\nabla_{\boldsymbol{w},\boldsymbol{p}}^2 l_i(P_{XY}^{(i)}, \boldsymbol{w}_{\mathrm{PM}i})\right]\operatorname{diag}\left\{P_{XY}^{(i)}\right\} \left[\nabla_{\boldsymbol{w},\boldsymbol{p}}^2 l_i(P_{XY}^{(i)}, \boldsymbol{w}_{\mathrm{PM}i})\right]^{\mathrm{T}} \\ &= 4\,\mathbb{E}_{P_{XY}^{(i)}}[\nabla_{\boldsymbol{w}}\boldsymbol{q}(X, Y; \boldsymbol{w}_{\mathrm{PM}i})\nabla_{\boldsymbol{w}}\boldsymbol{q}(X, Y; \boldsymbol{w}_{\mathrm{PM}i})^{\mathrm{T}}]. \end{aligned} \tag{56}$$

Finally,we have

$$R_i\left(\hat{\boldsymbol{f}}_{\mathrm{PM}i}, \hat{\boldsymbol{g}}_{\mathrm{PM}i}\right) = \frac{1}{n}\operatorname{rank}\left(\nabla_{\boldsymbol{w}}^2 L_i(\boldsymbol{w})\big|_{\boldsymbol{w}=\boldsymbol{w}_{\mathrm{PM}i}}\right) + O\left(\frac{1}{n^2}\right). \tag{57}$$

Thus Theorem 4.1 is proved.

REFERENCES

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[2] 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.

[3] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Exploiting shared representations for personalized federated learning," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 18–24 Jul 2021, pp. 2089–2099.

[4] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

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

[6] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[7] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[8] F. Grimberg, M.-A. Hartley, S. P. Karimireddy, and M. Jaggi, "Optimal model averaging: Towards personalized collaborative learning," *arXiv preprint arXiv:2110.12946*, 2021.

[9] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2021.

[10] Z. Charles and J. Konečnỳ, "Convergence and accuracy trade-offs in federated learning and meta-learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 2575–2583.

[11] W. Feller, *An introduction to probability theory and its applications, vol 2*. John Wiley & Sons, 2008.

[12] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. MIT press, 2009.

[13] A. Nilsson, S. Smith, G. Ulm, E. Gustavsson, and M. Jirstrand, "A performance evaluation of federated learning algorithms," in *Proceedings of the Second Workshop on Distributed Infrastructures for Deep Learning*, 2018, pp. 1–8.

[14] K. J. Rahman, F. Ahmed, N. Akhter, M. Hasan, R. Amin, K. E. Aziz, A. M. Islam, M. S. H. Mukta, and A. N. Islam, "Challenges, applications and design aspects of federated learning: A survey," *IEEE Access*, vol. 9, pp. 124 682–124 700, 2021.

[15] S.-L. Huang, A. Makur, G. W. Wornell, and L. Zheng, "On universal features for high-dimensional learning and inference," *arXiv preprint arXiv:1911.09105*, 2019.

[16] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.

[17] M. Thomas and A. T. Joy, "Elements of information theory," 2006.

[18] S.-L. Huang, X. Xu, L. Zheng, and G. W. Wornell, "An information theoretic interpretation to deep neural networks," in *2019 IEEE international symposium on information theory (ISIT)*. IEEE, 2019, pp. 1984–1988.

[19] I. Csiszár, "The method of types [information theory]," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2505–2523, 1998.

[20] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," 2004.

[21] G. W. Stewart, "Matrix perturbation theory," 1990.