Non-IID Federated Learning with Sharper Risk Bound

Bojian Wei, Jian Li*, Yong Liu, Weiping Wang

***Abstract*—In federated learning, the non-IID (not independently or identically distributed) data partitioning impairs the performance of the global model, which is a severe problem to be solved. Despite the extensive literature related to the algorithmic novelties and optimization analysis of federated learning, there has been relatively little theoretical research devoted to studying the generalization performance of non-IID federated learning. The generalization research of non-IID federated learning is still lack of effective tools and analytical approach. In this paper, we propose weighted local Rademacher complexity to pertinently analyze the generalization properties of non-IID federated learning and derive a sharper excess risk bound based on weighted local Rademacher complexity, where the convergence rate is much faster than the existing bounds. Based on the theoretical results, we present a general framework FedALRC to lower the excess risk without additional communication costs compared to some famous methods, such as FedAvg. Through extensive experiments, we show that FedALRC outperforms FedAvg, FedProx and FedNova, and those experimental results coincide with our theoretical findings.**

***Index Terms*—Federated learning, non-IID, excess risk bound, generalization analysis.**

I. Introduction

FEDERATED learning (FL) [1] is a new machine learning paradigm where a large number of clients collaboratively train a model under the coordination of a central server. In FL, the raw data of each client is stored locally, other clients and the central server have no access to it. Instead, the global model is updated by alternately performing local training and server aggregating. One main characteristic of FL is non-IID data partitioning across clients, such heterogeneity leads to the performance degradation and convergence slowdown.

To solve the problem, researchers have studied FL under non-IID setting through different approaches [2]–[6]. `FedAvg` [1] is the most common algorithm used to reduce communication costs by running multi-step SGD (Stochastic Gradient Descent) locally, and many works [7], [8] have proved its convergence by assuming the local iterations are the same. However, the performance of `FedAvg` drops to some extent in non-IID setting with different local iterations for the objective inconsistency [9]. `FedProx` [10] adds a proximal term to local objectives to constrain the gap between local models and the global model. Adaptive learning rates [11] have been extended to FL, which generalizes server momentum [12] without increasing communication costs. `FedNova` [9] is proposed to tackle the objective inconsistency problem and it could be combined with some acceleration techniques, while it only focus on the heterogeneity of local iterations. Other algorithms [13]–[16] are proposed to further reduce the impact of non-IID or accelerate the convergence, but they either require additional communication and memory [14] or only adapts to neural networks [13] according to the conclusions in [9].

Lack of Generalization Analysis. Although many related work has put forward different methods to solve the non-IID problem in FL, they are designed to improve the performance of FL from a specific aspect [17], [18], or to transfer some classic algorithms to FL, such as local momentum [19], variance reduction [14], [20], normalization [9] and adaptivity [11]. The relevant theoretical research is not suffcient, especially the generalization analysis. Most theoretical studies pay more attention to the convergence analysis of FL algorithm from the perspective of optimization [21], [22], where many works have analyzed federated optimization under homogeneity [23] or heterogeneity [23], [24] setting and tried to explore the convergence under milder assumptions. Only a few work give the generalization bound for FL. Agnostic federated learning [25], [26] provides a new point on classical FL, but the target is to optimize the worst case in the hypothesis space, where the theoretical analyses are overly pessimistic and the algorithm often performs not well in practice. And, they give a generalization bound with the convergence rate of $\mathcal{O}(\sqrt{\frac{\log n}{n}})$ for binary-classification, where n is the total number of samples among all the clients. Considering the distribution discrepancy, three approaches [27] are introduced to improve the performance of personalized FL, including hypothesis-based clustering, data interpolation and model interpolation. However, the convergence rate of their generalization bound is $\mathcal{O}(\frac{K}{\sqrt{n}})$, where K is the number of clients. Thus, it is necessary to study how to reduce the generalization error of FL with faster convergence rate.

In this paper, we present a novel generalization analysis for non-IID FL and derive a sharper excess risk bound based on weighted local Rademacher complexity, where the convergence rate meets the current results in centralized learning and the theory is a non-trivial extension of the existing bounds. Based on the theoretical analysis, we devise an effective algorithm to improve the performance of non-IID FL, which introduces a regularization scheme to constrain the

B. Wei is from Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, and also from School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China (e-mail: szwboj@126.com).

J. Li and W. Wang are from Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China (e-mail: lijian9026@iie.ac.cn; wangweiping@iie.ac.cn).

Y. Liu is from Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China (e-mail:liuyongsai@ruc.edu.cn).

*Jian Li is the corresponding author.

local Rademacher complexity. Experimental results validate the effectiveness of our theory and proposed algorithm.

Contributions:

- To the best of our knowledge, we propose weighted local Rademacher complexity to tackle the non-IID problem in FL for the first time and present the first generalization analysis of non-IID FL based on it. We derive a novel excess risk bound for non-IID FL with generalized linear models, including linear models, kernel methods, and so on, which is much sharper than the existing results ($\mathcal{O}(\frac{K}{\sqrt{n}})$) with the convergence rate of $\mathcal{O}(\frac{K}{n})$ for shallow models.
- Motivated by our theory, we propose `FedALRC`, a general framework for non-IID FL to improve the generalization performance by constraining weighted local Rademacher complexity. Moreover, `FedALRC` also preserves fast convergence and low communication costs.
- Through extensive numerical experiments on various datasets, `FedALRC` outperforms other FL algorithms under the same non-IID setting. The effectiveness of weighted local Rademacher complexity is also valided by combining `FedALRC` with adaptive optimization.

A. Related Work

Recently, several work studies the generalization performance of FL under non-IID setting. Agnostic federated learning [25] regards the global distribution on the server as the mixture of local distributions, and ensures the fairness with minimax framework. SCAFFOLD [14] reduces the generalization error of non-IID FL by constraining the variance among local models. Based on sketch and differential privacy [28], an efficient approach for cross-silo federated learning to rank [29] provides a privacy-preserving strategy for frequency query and bounds the related estimation error. FedGen [30] is a data-free FL algorithm based on knowledge distillation, which trains a generator on the server to gather the information from clients. The authors also give a generalization bound for FedGen, where the convergence rate is $\mathcal{O}(\sqrt{\frac{K}{n}})$. Motivated by the margin-based generalization bound, DMFL [31] introduces a dynamic constraint on the local objective to allocate bigger margin to the class with less samples.

Excess risk bounds for non-IID FL. As mentioned above, the existing bounds still have much room for improvement. The existing generalization bounds for non-IID FL are mainly based on Rademacher complexity [25], [30], [32], where they only consider the worst case in hypothesis space and the convergence rate can not be better than $\mathcal{O}(\frac{K}{\sqrt{n}})$ or $\mathcal{O}(\sqrt{\frac{K}{n}})$. According to [33], the excess risk of a learning algorithm is twice the generalization error, so the present excess risk bounds are of the same order as the corresponding generalization bounds. Thus, improving the excess risk bound can provide a stronger guarantee for non-IID FL, which is of great significance to improve its generalization performance.

Local Rademacher complexity. In recent years, many researchers have applied local Rademacher complexity [34] in centralized learning to obtain better generalization bounds [35]–[39]. However, how to use local Rademacher complexity to non-IID FL to derive a sharper bound is still an open problem. A naive way is to directly convert the excess risk into the weighted sum of the local excess risk, but this will lead to a loose bound and does not conform to the mechanism of FL. Specifically, the former is the weighted sum of the optimal solutions for local objectives, while the latter is the optimal solution for the weighted sum of local objectives (global optima). To this end, we introduce a weighted counterpart of local Rademacher complexity to analyze non-IID FL and derive a sharper excess risk bound, which is consistent with the global objective of non-IID FL.

TABLE I
CRITICAL NOTATIONS.

Notation	Interpretation
$\mathcal{X}$, $\mathcal{Y}$	input space, label space
$\mathcal{L}$, $\mathcal{H}$	loss space, hypothesis space
f, $\ell_f(\mathbf{x}, y)$	labeling function, loss function
$\mathcal{H}_r$, $\mathcal{L}^*$	localized hypothesis space, excess loss space
$P\ell_f$, $P_n\ell_f$	expected loss of FL, empirical loss of FL
$(\mathbf{x}^k, y^k)$	training samples on the k-th client
$\widehat{\mathcal{R}}(\mathcal{L}, p)$	empirical weighted Rademacher complexity of $\mathcal{L}$
$\mathcal{R}(\mathcal{L}_r, p)$	expected weighted local Rademacher complexity of $\mathcal{L}$
$\widehat{\mathcal{R}}(\mathcal{H}_r, p)$	empirical weighted local Rademacher complexity of $\mathcal{H}_r$
$\mathcal{R}(\mathcal{L}_r^*, p)$	expected weighted local Rademacher complexity of $\mathcal{L}^*$
r^*	fixed point of $\mathcal{R}(\mathcal{L}_r^*, p)$
$\mathbf{W}$, $\phi(\mathbf{x})$	classifier, feature mapping
$\{\lambda_j^k\}$, $\{\epsilon_i^k\}$	singular values, Rademacher variables

The rest of this paper is organized as follows. In Section II, we give the general notations and definitions used in this paper. We present the theoretical results and some critical discussions in Section III. In Section IV, we propose the two counterparts of `FedALRC` and the extension to adaptive learning rates, we also demonstrate the local computation and communication cost. Extensive experiments are illustrated with explanations in Section V. We conclude in Section VI and give the proofs in Appendix.

II. PRELIMINARIES AND NOTATIONS

We mainly focus on the cross-silo non-IID FL setting, where all the clients participate in the training process per round. There are some general notations used in this paper and the critical symbols are listed in TABLE I.

Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the input space, $\mathcal{Y} \subseteq \mathbb{R}^C$ denote the label space, and $\mathcal{H}$ be the hypothesis space consisting of labeling functions $f : \mathcal{X} \to \mathcal{Y}$. In FL, there are K clients and a central server, where samples $(\mathbf{x}^k, y^k)$ on the k-th client is drawn i.i.d. from local distribution ρ_k with size of n_k, and data on different clients may not have the same distribution ($\rho_i \neq \rho_j$).

Let $\mathcal{L}$ be the family of loss functions associated to $\mathcal{H}$. Without loss of generality, we assume that the loss function $\ell_f(\mathbf{x}, y) = \ell(f(\mathbf{x}), y)$ on each client is bounded by B ($B > 0$). We denote by $P\ell_f$ the expected loss of FL:

$$P\ell_f = \sum_{k=1}^{K} p_k P^k \ell_f = \sum_{k=1}^{K} p_k \mathbb{E}_{(\mathbf{x}^k, y^k)\sim\rho_k}[\ell_f(\mathbf{x}^k, y^k)],$$

and $P_n\ell_f$ the corresponding empirical loss:

$$P_n\ell_f = \sum_{k=1}^{K} p_k P_n^k \ell_f = \sum_{k=1}^{K} p_k \frac{1}{n_k} \sum_{i=1}^{n_k} \ell_f(\mathbf{x}_i^k, y_i^k),$$

where $p_k = \frac{n_k}{n}$ ($n = \sum_{k=1}^{K} n_k$) is the aggregation weight.

Note that the global objective is the weigted sum of local objectives, and each local objective is unique related to non-IID setting. Thus, the traditional Rademacher complexity will no longer apply to this situation. We use the weighted counterpart to deal with the non-IID problem contrapuntally.

Definition 1 (Weighted Rademacher Complexity): Let $\mathcal{L}$ be the family of loss functions defined above, the empirical weighted Rademacher complexity of $\mathcal{L}$ is

$$\widehat{\mathcal{R}}(\mathcal{L}, p) = \mathbb{E}_\epsilon \left[\sup_{\ell_f \in \mathcal{L}} \sum_{k=1}^{K} \frac{p_k}{n_k} \sum_{i=1}^{n_k} \epsilon_i^k \ell_f(\mathbf{x}_i^k, y_i^k) \right],$$

where $\{(\mathbf{x}_i^k, y_i^k), ..., (\mathbf{x}_i^k, y_i^k)\}$ is a sample of size n_k on the k-th client, ϵ_i^ks are independent Rademacher variables, which are uniformly sampled from $\{-1, +1\}$. The weighted Rademacher complexity of $\mathcal{L}$ is $\mathcal{R}(\mathcal{L}, p) = \mathbb{E}[\widehat{\mathcal{R}}(\mathcal{L}, p)]$.

Though weighted Rademacher complexity is already applicable to non-IID FL setting, it ignores the fact that, the hypotheses selected by a learning algorithm belong to a subfamily in the hypothesis space with good performance. Thus, we will use weighted local Rademacher complexity to obtain a sharper bound.

Definition 2 (Weighted Local Rademacher Complexity): For any $r > 0$, the weighted local Rademacher complexity of $\mathcal{L}$ is defined as

$$\mathcal{R}(\mathcal{L}_r, p) = \mathcal{R}\left\{\ell_f | \ell_f \in \mathcal{L}, P\ell_f^2 \leq r\right\}.$$

Then, we get the corresponding localized hypothesis space:

$$\mathcal{H}_r := \{f | f \in \mathcal{H}, P\ell_f^2 \leq r\},$$

and we define the empirical weighted local Rademacher complexity of $\mathcal{H}_r$ as

$$\widehat{\mathcal{R}}(\mathcal{H}_r, p) = \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{H}_r} \sum_{k=1}^{K} \frac{p_k}{n_k} \sum_{i=1}^{n_k} \sum_{c=1}^{C} \epsilon_{ic}^k f_c(\mathbf{x}_i^k) \right],$$

where $f_c(\mathbf{x}_i^k)$ is the c-th value of $f(\mathbf{x}_i^k)$ w.r.t. the c-th class and ϵ_{ic}^ks are independent Rademacher variables, which are uniformly sampled from $\{-1, +1\}$.

In the following, we assume that ℓ_f is L-lipschitz for $\mathbb{R}^C$ equipped with the 2-norm, that is $|\ell_f(\mathbf{x}, y) - \ell_{f'}(\mathbf{x}, y)| \leq L\|f(\mathbf{x}) - f'(\mathbf{x})\|_2$, where many loss functions meet this condition, such as hinge loss, margin loss and their variants. And, this is also a commonly used assumption in generalization analysis for vector-valued labeling functions.

III. Sharper Risk Bound

In this section, we present a sharper excess risk bound for non-IID FL. All the proofs can be found in the appendix.

We introduce the following excess loss space:

$$\mathcal{L}^* := \{\ell_f - \ell_{f^*} | \ell_f \in \mathcal{L}\},$$

where f^* denotes the labeling function which satisfies $\ell_{f^*} = \inf_{f \in \mathcal{H}} P\ell_f$, Then, we define the weighted local Rademacher complexity of $\mathcal{L}^*$ as

$$\mathcal{R}(\mathcal{L}_r^*, p) = \mathcal{R}\{\ell_f - \ell_{f^*} | \ell_f \in \mathcal{L}, P(\ell_f - \ell_{f^*})^2 \leq r\}.$$

Theorem 1 (Excess Risk Bound): Let $\widehat{f}$ be the labeling function satisfying $\ell_{\widehat{f}} = \inf_{f \in \mathcal{H}} P_n\ell_f$. Then, for any $\delta \in (0, 1)$, $\forall f \in \mathcal{H}_r$ and $\forall G > 1$, with probability at least $1 - \delta$, the following bound holds:

$$P(\ell_{\widehat{f}} - \ell_{f^*}) \leq \frac{800G}{B} r^* + \frac{(16G + 12)B \log(1/\delta)}{n}, \quad (1)$$

where r^* is the fixed point of $\mathcal{R}(\mathcal{L}_r^*, p)$, which denotes the unique positive solution of $\mathcal{R}(\mathcal{L}_r^*, p) = r$.

Discussion. By weighted local Rademacher complexity $\mathcal{R}(\mathcal{L}_r^*, p)$, a smaller class $\mathcal{L}_r^* \subseteq \mathcal{L}^*$ with small variance around the optimal hypothesis is selected, measuring by a fixed radius r. Since Rademacher complexity only considers the worst case of the hypothesis space, the previous bounds can not converge faster than $\mathcal{O}(1/\sqrt{n})$, while our excess risk bound mainly depends on the fixed point r^*. Note that the rate of r^* can not be worse than the rate obtained by Rademacher complexity, which means that our bound is at least on the same level as [27]. For instance, in centralized kernel learning, the rate of r^* can achieve up to $\mathcal{O}(\frac{1}{n})$ for linear kernels, polynomial kernels and Gaussian kernels. Thus, our bound can be much better and we give a specific demonstration as follows.

Consider that $\mathcal{H} := \{f | f = \mathbf{W}^T\phi(\mathbf{x}), \|\mathbf{W}\| \leq 1\}$, where $\phi(\cdot) \subseteq \mathbb{R}^D$ denotes a fixed feature mapping, we denote by $\mathbf{W}_k$ the learnable parameters on the k-th client. In the following theorem, we give an estimate of weighted local Rademacher complexity and a sharper risk bound.

Theorem 2: Assume that $\mathbb{E}[\phi(\mathbf{x}^k)^T\phi(\mathbf{x}^k)] \leq 1$ and $\|\mathbf{W}_k\| \leq 1$ for client k. Let $\mathbf{W}_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T$ be the SVD decomposition of $\mathbf{W}_k$, where $\mathbf{U}_k \in \mathbb{R}^{D\times D}$ and $\mathbf{V}_k \in \mathbb{R}^{C\times C}$ are unitary matrices, and $\mathbf{\Sigma}_k \in \mathbb{R}^{D\times C}$ is diagonal with singular values $\{\lambda_j^k\}$ in descending order. Then, for any $r > 0$, we have

$$\mathcal{R}(\mathcal{L}_r^*, p) \leq \inf_{\vartheta \geq 0} \sum_{k=1}^{K} p_k \left(\sqrt{\frac{2\vartheta r}{n_k}} + 2\sqrt{2}L \frac{\sum_{j>\vartheta} \lambda_j^k}{\sqrt{n_k}} \right),$$

where $\vartheta \in \mathbb{N}$ is the truncated threshold. For any $\delta \in (0, 1)$ and $\vartheta \geq 0$, $\forall f \in \mathcal{H}$, with probability at least $1 - \delta$, we have

$$\begin{aligned} &P(\ell_{\widehat{f}} - \ell_{f^*}) \\ \leq &\mathcal{O}\left(\frac{K\vartheta + \sum_{k=1}^{K} \sqrt{n_k} \sum_{j>\vartheta} \lambda_j^k}{n} + \frac{\log(1/\delta)}{n} \right), \end{aligned}$$

where $\mathcal{O}$ swallows all constants (including G, L and B).

The above theorems demonstrate that the excess risk of non-IID FL is determined by the weighted local Rademacher complexity $\mathcal{R}(\mathcal{L}_r^*, p)$, and $\mathcal{R}(\mathcal{L}_r^*, p)$ is determined by the sum of the tail sum of $\mathbf{W}_k$'s singular values. This inspires us to reduce the excess risk by constraining the sum of the tail sum of $\mathbf{W}_k$'s singular values, so as to improve the generalization performance of non-IID FL. We make the following discussions:

- In the worst case ($\vartheta = 0$), weighted local Rademacher complexity degrades into weighted Rademacher complexity, and we get a convergence rate of $\mathcal{O}(\frac{\sum_{k=1}^{K}\sqrt{n_k}}{n})$. If data on each client has the same size, then the convergence rate is $\mathcal{O}(\sqrt{\frac{K}{n}})$, which is already better than $\mathcal{O}(\frac{K}{\sqrt{n}})$ [27].
- When $\mathbf{W}_k$ has a finite rank ϑ such that its singular values satisfy $\lambda_j^k = 0$ for all $j > \vartheta$, which means that the tail sum of singular values is zero. Thus, the rate of the fixed point is inversely proportional to the number of samples n_k: $r^* = \mathcal{O}(\frac{K\vartheta}{n})$. Then, the convergence rate of the excess risk bound is $\mathcal{O}(\frac{K\vartheta}{n})$, which is much sharper than the previous results.
- When the singular values of $\mathbf{W}_k$ decay exponentially, that is $\sum_{j>\vartheta}\lambda_j = \mathcal{O}(e^{-\vartheta})$, then it holds $r^* = \mathcal{O}(\frac{K\log n}{n})$, and we also obtain a faster convergence rate $\mathcal{O}(\frac{K}{n})$.

***Remark** 1 (Beyond horizontal federated learning):* In this paper, by restricting the complexity of hypothesis space, we derive sharper excess risk bounds for the horizontal federated learning scenarios where the local clients share the same feature space but differ in samples [40]. Nevertheless, the techniques presented here can be extended to vertical federated learning methods that share the same sample ID space but with different feature spaces across local clients. Specifically, one should carefully define the (local) Rademacher complexity for vertical FL settings and then impose a constraint on local Rademacher complexity to guarantee a smaller hypothesis space around the target function [34]. This also motivates an improved algorithm for the vertical FL methods that minimize the training loss and reduces the empirical local Rademacher complexity at the same time.

***Remark** 2 (Novelty):* Both $\mathbf{W}_k$ with finite rank and $\mathbf{W}_k$ with exponentially decaying singular values have a fixed point r^* that mainly depends on $\mathcal{O}(\frac{K}{n})$. In these cases, the excess risk bound for FL achieves a linear dependence on the total sample size and is independent of the number of classes C. Compared to the existing results based on weighted Rademacher complexity [25], [27], where the related bounds can not converge faster than $\mathcal{O}(K/\sqrt{n})$, we obtain a sharper risk bound based on weighted local Rademacher complexity, which provides a stronger generalization guarantee for FL with non-IID data.

***Remark** 3 (Generality):* Note that the form of $\mathbf{W}^T\phi(\mathbf{x})$ contains various learning algorithms. When $\phi(\mathbf{x}) = \mathbf{x}$, then $\mathcal{H}$ becomes a linear space, and we can always obtain a sharper risk bound for all linear models. When $\phi(\cdot)$ is non-linear, $\mathcal{H}$ can represent many different hypotheses, including generalized linear models, kernel methods, random fourier features, shallow neural networks and pre-trained models with fine-tuning. Moreover, Theorem 2 can be applied to other learning tasks, such as multi-objective learning and distributed learning. Thus, our excess risk bound has a wide range of applicability, which can provide better theoretical guarantee for lots of mainstream algorithms.

***Remark** 4 (Proof Novelty):* According to the definition of weighted local Rademacher complexity, it is consistent with the global objective of non-IID FL, where the $\sup$ operation is outside the weighted sum operation, and there is no existing method that can be directly used to derive the excess risk bound for non-IID FL based on weighted local Rademacher complexity. To get the sharper risk bound, we first propose two propositions for weighted local Rademacher complexity and give the related proofs, then we derive an error bound for a generalized loss space with the propositions and Talagrand's inequality, and finally extend it to the excess loss space to get the final result. Thus, our theoretical finding is a non-trivial extension of the existing results.

Algorithm 1 `FedALRC` for shallow models. ν is the fraction of participation, $\mathbf{w}$ is the model parameters, ϖ_k is the local iterations and $\mathcal{B}$ is the mini-batch size.

Server Aggregating:
  Initialize $\mathbf{w}^0$
  for each communication round $t = 1, 2, ..., T$ **do**
    $m_\nu = \max(\nu m, 1)$
    $S_t :=$ random set of m_ν clients
    for client $k \in S_t$ **in parallel do**
      $\Delta_k^{t-1} \leftarrow$ Local Training$(k, \mathbf{w}^{t-1})$
    end for
    $\Delta^{t-1} = \sum_{k=1}^{K} p_k \Delta_k^{t-1}$
    $\mathbf{w}^t = \mathbf{w}^{t-1} - \eta\Delta^{t-1}$
  end for
Local Training$(k, \mathbf{w}^{t-1})$**:**
  $\mathbf{w}_k^{t-1} = \mathbf{w}^{t-1}$
  for epoch$= 1, ..., E$ **do**
    for each batch $(\mathbf{x}_i^k, y_i^k)_{i=1,...,\mathcal{B}}$ **do**
      $\mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T = \mathbf{w}_k^{t-1} - \eta_l \nabla_{\mathbf{w}_k^{t-1}} \frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} \ell_f(\mathbf{x}_i^k, y_i^k)$
      $\mathbf{w}_k^{t-1} = \mathbf{U}_k\mathbf{\Sigma}_k^{\vartheta, \alpha\eta_l}\mathbf{V}_k^T$
    end for
  end for
  $\Delta_k^{t-1} = \frac{\mathbf{w}^{t-1} - \mathbf{w}_k^{t-1}}{\varpi_k}$

IV. Algorithm

In this section, we present a general framework `FedALRC` to minimize weighted local Rademacher complexity during FL training based on our sharper excess risk bound.

A. Federated Averaging with Local Rademacher Complexity

According to Theorem 1, the excess risk can be lowered by reducing weighted local Rademacher complexity. Thus, a simple approach is to add weighted local Rademacher complexity to the global objective function as a regularization term. However, we can not estimate $\mathcal{R}(\mathcal{L}_r^*, p)$ directly on the server side, because it is data-dependent and data is not interactive under FL setting.

Note that $\mathcal{R}(\mathcal{L}_r^*, p) \leq \sum_{k=1}^{K} p_k \mathcal{R}^k(\mathcal{L}_r^*)$, where $\mathcal{R}^k(\mathcal{L}_r^*)$ denotes the local Rademacher complexity of $\mathcal{L}_r^*$ on the k-th client, we know that $\mathcal{R}(\mathcal{L}_r^*, p)$ will decrease with the decrease of $\mathcal{R}^k(\mathcal{L}_r^*)$. Therefore, we impose the constraint of $\mathcal{R}(\mathcal{L}_r^*, p)$ to the local objective function on each client to improve generalization performance.

For shallow models: According to Theorem 2, weighted local Rademacher complexity is bounded by the sum of the

tail sum of $\mathbf{W}_k$'s singular values. Thus, the local objective function on the k-th client is formed as

$$\arg\min_{f\in\mathcal{H}_r^*} \frac{1}{n_k}\sum_{i=1}^{n_k}\ell_f(\mathbf{x}_i^k, y_i^k) + \alpha\sum_{j>\vartheta}\lambda_j^k,$$

where α is a tunable parameter.

The minimization of the sum of partial singular values can be difficult to implement, and thus we change it into a two-step optimization [37] in Algorithm 1. First, the local model $\mathbf{W}_k$ is updated through SGD w.r.t. the empirical loss except for the tail sum of singular values. Second, $\mathbf{W}_k$ is updated by singular value thresholding $\mathbf{W}_k = \mathbf{U}_k\mathbf{\Sigma}_k^{\vartheta,\alpha\eta_l}\mathbf{V}_k^T$, where $\mathbf{\Sigma}_k^{\vartheta,\alpha\eta_l}$ is diagonal with

$$[\mathbf{\Sigma}_k^{\vartheta,\alpha\eta_l}]_{jj} = \begin{cases} \max(0, [\mathbf{\Sigma}_k]_{jj} - \alpha\eta_l) & j \le \vartheta, \\ [\mathbf{\Sigma}_k]_{jj} & j > \vartheta. \end{cases}$$

For deep models: The updating process in deep models can be very complex, and the related characteristics will be lost by directly applying Theorem 2. Thus, we propose an empirical method to minimize the empirical weighted local Rademacher complexity $\widehat{\mathcal{R}}(\mathcal{L}_r^*, p)$. For the k-th client, we sample Q times Rademacher variables $\{\epsilon_{ic}^k\}_{i=1,\dots,n_k}^{c=1,\dots,C}$ and then calculate the empirical local Rademacher complexity by $\Re^k = \frac{1}{n_k C}\sum_{i=1}^{n_k}\sum_{c=1}^{C}\epsilon_{ic}^k f_c(\mathbf{x}_i^k)$ such that $P(\ell_f - \ell_{f^*})^2 \le r$. Taking the average across Q times, we add the average $\Re^k$ to the local objective function as a regularization term, so the k-th local objective function is formed as

$$\arg\min_{f\in\mathcal{H}_r^*} \frac{1}{n_k}\sum_{i=1}^{n_k}\ell_f(\mathbf{x}_i^k, y_i^k) + \alpha\Re^k. \tag{2}$$

Note that the server aggregating process is the same, so we give the pseudo-code of local training in Algorithm 2.

As for the optimization in non-IID FL, the objective inconsistency [9] has been proposed, which shows that `FedAvg` can not converge to the global objective $P_n\ell_f$. To this end, we correct each local update Δ_k by dividing the corresponding local iterations ϖ_k when SGD is applied to local training. Further, `FedAvg` can be accelerated by two-sided learning rates [41]. Thus, we introduce a server-side learning rate η in server aggregating to reduce the convergence slowdown.

Combination with adaptive optimization. The server aggregating algorithm used in the above experiments is pseudo-gradient descent, while some techniques [11], [19] have been proposed to improve the performances of FL algorithms on the server side. Thus, we combine adaptive learning rates (general server momentum) in Algorithm 3, named `Ada-FedALRC`, where the local training process is the same as Algorithm 1 for shallow models and Algorithm 2 for deep models.

Computation of local training. `FedALRC` only introduces an additional regularization term in the local objective, which is not much more complicated compared to FedAvg. For shallow models, the additional complexity of local training is the SVD decomposition. If the dimension of $\mathbf{W}$ is very high, `FedALRC` will bring extra computing costs that can not be ingored. But, the additional costs can be accepted when the dimension of $\mathbf{W}$ is mild. For deep models, the additional computation comes from the estimation of local Rademacher complexity, where we sample a group of Rademacher variables $\{\epsilon_{ic}^k\}_{i=1,\dots,\mathcal{B}}^{c=1,\dots,C}$, which can be treated as a $B \times C$ matrix. The Q-times average for $\Re^k$ is an inplace operation, so the extra memory remains the $B \times C$ matrix. Note that the regularization term in FedProx is $\|\mathbf{w} - \mathbf{w}_{global}\|^2$, so our computing costs may not greater than FedProx's. Above all, although `FedALRC` does introduce additional computation in local training, the magnitude is not very large compared with other algorithms, so it can be widely used in current devices.

Communication cost. The communication cost of FL is mainly caused by the interaction of model parameters (or gradients) between clients and server during training process. `FedALRC` and `Ada-FedALRC` leave the optimization of local Rademacher complexity on client-side, which interact the same objects (model parameters) pre round as `FedAvg` and `FedProx`. Besides, we show that `FedALRC` and `Ada-FedALRC` converge faster than other algorithms in next section. Thus, our proposed algorithms have no additional communication cost compared with `FedAvg` and `FedProx`.

***Remark** 5 (Applications):* According to Remark 3, our theory fits many kinds of models. And, we propose two counterparts of `FedALRC` related to the type of models.

Algorithm 2 `FedALRC` for deep models. $\mathbf{w}$ is the model parameters, ϖ_k is the local iterations and $\mathcal{B}$ is the mini-batch size. The procedure of server aggregating is the same as that in Algorithm 1.

Local Training$(k, \mathbf{w}^{t-1})$**:**
$\mathbf{w}_k^{t-1} = \mathbf{w}^{t-1}$
for epoch$= 1, ..., E$ **do**
  for each batch $(\mathbf{x}_i^k, y_i^k)_{i=1,\dots,\mathcal{B}}$ **do**
    $\Re^k = 0$
    if $P(\ell_f - \ell_{f^*})^2 \le r$ **then**
      for $q = 1, ..., Q$ **do**
        Sample Rademacher variables $\{\epsilon_{ic}^k\}_{i=1,\dots,\mathcal{B}}^{c=1,\dots,C}$
        $\Re^k \leftarrow \Re^k + |\frac{1}{\mathcal{B}C}\sum_{i=1}^{\mathcal{B}}\sum_{c=1}^{C}\epsilon_{ic}^k f_c(\mathbf{x}_i^k)|$
      end for
      $\Re^k = \Re^k / Q$
    end if
    $\mathfrak{L} = \frac{1}{\mathcal{B}}\sum_{i=1}^{\mathcal{B}}\ell_f(\mathbf{x}_i^k, y_i^k) + \alpha\Re^k$
    $\mathbf{w}_k^{t-1} = \mathbf{w}_k^{t-1} - \eta_l\nabla_{\mathbf{w}_k^{t-1}}\mathfrak{L}$
  end for
end for
$\Delta_k^{t-1} = \frac{\mathbf{w}^{t-1} - \mathbf{w}_k^{t-1}}{\varpi_k}$

Algorithm 3 Server aggregating of `Ada-FedALRC`.

Initialize $\mathbf{w}^0$ and $v^0 = \text{eps}$
for each communication round $t = 1, 2, ..., T$ **do**
  $m_\nu = \max(\nu m, 1)$
  $S_t :=$ random set of m_ν clients
  for client $k \in S_t$ **in parallel do**
    $\Delta_k^{t-1} \leftarrow$ Local Training$(k, \mathbf{w}^{t-1})$
  end for
  $\Delta^{t-1} = \eta\sum_{k=1}^{K} p_k\Delta_k^{t-1}$
  $v^t = v^{t-1} + (\Delta^{t-1})^2$
  $\mathbf{w}^t = \mathbf{w}^{t-1} - \beta_\eta\frac{\Delta^{t-1}}{\sqrt{v^t}+\text{eps}}$
end for

TABLE II
STATISTICAL INFORMATION OF DATASETS.

Datasets	Traing Size	Testing Size	Feature Dimension	Number of Classes
usps	7291	2007	256	10
pendigits	7494	3498	16	10
satimage	4435	2000	36	6
letter	15000	5000	16	26
vowel	528	462	10	11
MNIST	60000	10000	28×28	10
CIFAR-10	50000	10000	3×32×32	10

Based on the definition of local Rademacher complexities, our theoretical finding aims at the supervised federated learning scenario, especially the multi-classification with vector-value outputs. Therefore, our proposed scheme can be applied to non-IID FL with vector-value classification tasks.

V. EXPERIMENTS

In this section, we evaluate all algorithms on various real-world datasets[1] with non-IID partitioning.

A. Experimental Setup

We train a linear model on several LIBSVM [42] datasets, a LeNet network on *MNIST* and a VGG-11 network on *CIFAR-10*, where all the train sets are partitioned across 20 clients (16 clients for CIFAR-10) using a Dirichlet distribution $\texttt{Dir}_K(0.1)$ [13] and the original test set of each dataset is used to evaluate the performance of the global model. The statistical information of LIBSVM datasets is listed in TABLE II. The client-side learning rate η_l is decayed in the same way as [9]. To ensure the fairness of comparison, we tune η_l for `FedAvg` and apply the same value to the other algorithms. We run each experiment with 3 random seeds and record the average and standard deviation. All clients perform $E = 2$ local epochs in the following experiments. All the experiments are conducted on a Linux server equipped with one NVIDIA GeForce 2080ti, and all the algorithms are implemented by Pytorch.

B. Experiments with Non-IID Partitioning

In TABLE III, we compare the performance of `FedALRC` and three mainstream algorithms (`FedAvg`, `FedProx` and `FedNova`) on various datasets with non-IID data partitioning. The characteristics of these algorithms are summarized in TABLE IV. We run each experiment for 100 rounds and apply SGD to local training. We fix the mini-batch size per client as 64 for MNIST, 32 for CIFAR-10 and 16 for the rest of datasets. The client-side learning rate η_l is tuned from $\{0.001, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 0.7\}$ for `FedAvg`, the scale parameter γ of server-side learning rate η is tuned from $\{0.9, 1.0, 1.2, 1.5, 2.0, 2.5\}$ for `FedALRC` and the proximal parameter μ_{px} for `FedProx` is tuned from $\{0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1\}$.

[1]https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

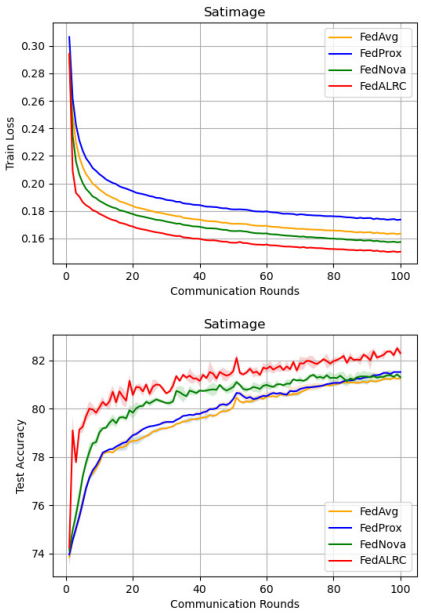

Fig. 1. Results on *Satimage* under Non-IID Setting.

For linear model, we manually tune the SVD threshold ϑ from $\{1, ..., \min(D, C)\}$ and the regularization parameter α from $\{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 100\} \times 10^{-3}$. For DNN (Deep Neural Network), we set $r = 1$ to constrain $\mathcal{H}_r^*$, where the minimum loss ℓ_f^* can be reguarded as 0 for simplicity, and we introduce weight decay ($\times 10^{-4}$) into local training to avoid over-fitting of deep network.

According to the results in TABLE III, we can know that `FedALRC` is significantly better than other algorithms with confidence at level 95% on most datasets. By modifying server-side learning rate and constraining local Rademacher complexity, `FedALRC` yields a significant improvement up to 8% compared to `FedAvg` and `FedProx`, and outperforms `FedNova` with a clear margin. In Fig. 1, we observe that `FedALRC` not only achieves the best performance under non-IID setting, but also converges faster than the other algorithms, which is consistent with our theory. There is an interesting phenonmenon that the hovering amplitude of `FedALRC` is larger than other methods. This indicates that `FedALRC` may be sensitive to the server-side learning rate.

The reason why the improvements of `FedALRC` on some datasets is not significant compared with other algorithms is that our algorithm aims to reduce the excess risk, in other words, the generalization error. And, the excess risk is measured by the weighted local Rademacher complexity, which is data-dependent. Thus, the performance may be different when we train different models with different datasets, especially when models or datasets do not fit the assumptions perfectly.

We also compare the performance of `Ada-FedALRC` with `FedAdagrad` [11]. The local training process of `FedAdagrad` is the same as `FedAvg`. The threshold eps is set as 10^{-3} and β_η is tuned from $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.5\}$.

From the right side of TABLE III, we can know that `Ada-FedALRC` generally outperforms `FedAdagrad`, which

TABLE III
TEST ACCURACY (%) OF DIFFERENT FL ALGORITHMS ON DIFFERENT DATASETS WITH NON-IID DATA PARTITIONING. WE BOLD THE NUMBERS OF THE BEST METHOD AND UNDERLINE THE NUMBERS OF OTHER METHODS WHICH ARE NOT SIGNIFICANTLY WORSE THAN THE BEST ONE.

Datasets	FL Algorithms				Adaptive FL Algorithms	
	FedAvg	FedProx	FedNova	FedALRC	FedAdagrad	Ada-FedALRC
usps	82.41±0.07	82.36±0.12	82.93±0.08	**83.03±0.12**	82.63±0.02	**82.98±0.02**
pendigits	77.47±0.75	77.98±0.60	78.74±0.49	**79.20±0.14**	77.78±0.63	**77.90±0.37**
satimage	81.27±0.06	81.52±0.05	81.30±0.15	**82.30±0.18**	82.12±0.26	**82.78±0.31**
letter	29.90±0.03	29.98±0.00	29.69±0.05	**30.04±0.02**	30.01±0.06	**30.26±0.03**
vowel	35.43±0.37	35.57±0.27	37.81±0.10	**38.24±0.10**	36.80±0.47	**37.45±0.61**
MNIST	97.49±0.04	97.42±0.05	97.24±0.14	**98.15±0.13**	97.97±0.08	**98.15±0.13**
CIFAR-10	63.70±1.02	63.55±1.05	71.85±0.33	**72.28±0.23**	68.46±1.22	**71.63±0.97**

TABLE IV
CHARACTERISTICS OF FL ALGORITHMS.

Algorithms	Proximal Term	Local Update	Server-Side Learning Rate
FedAvg	✗	$\mathbf{w}^t - \mathbf{w}_k^t$	1
FedProx	✓	$\mathbf{w}^t - \mathbf{w}_k^t$	1
FedNova	✗	$\frac{\mathbf{w}^t - \mathbf{w}_k^t}{\varpi_k}$	$\sum_{k=1}^K p_k \varpi_k$
FedALRC	✗	$\frac{\mathbf{w}^t - \mathbf{w}_k^t}{\varpi_k}$	$\gamma \sum_{k=1}^K p_k \varpi_k$

TABLE V
TEST ACCURACY (%) OF FEDALRC WITH DIFFERENT REGULARIZERS.

Algorithms	Datasets			
	usps	*pendigits*	*satimage*	*letter*
FedALRC-g	82.36±0.16	78.33±0.39	81.38±0.19	29.94±0.03
FedALRC-l	**83.03±0.12**	**79.20±0.14**	**82.30±0.18**	**30.04±0.02**

means that our method combined with adaptive algorithm can also improve the generalization capacity. Moreover, it should be noted that the test accuracy of Ada-FedALRC is lower than that of FedALRC on some datasets. The reason for this also refers to the Rademacher complexity-style bound, we think that the Adagrad-style algorithm may not be able to guarantee a lower generalization error on these datasets, which affects the improvement of the related generalization performance.

C. Comparison with Weighted Rademacher Complexity

The key point of FedALRC is introducing a regularization term based on weighted local Rademacher complexity to reduce the excess risk. According to [32], a regularization term based global Rademacher complexity is designed to reduce the excess risk of non-IID FL. The empirical global Rademacher complexity for k-th device $\Re^k$ can be formed as

$$\Re^k = \begin{cases} \mathrm{Tr}(\mathbf{W}_k), & \text{shallow models,} \\ \frac{1}{n_k C} \sum_{i=1}^{n_k} \sum_{c=1}^{C} \epsilon_{ic}^k f_c(\mathbf{x}_i^k), & \text{deep models.} \end{cases}$$

Here, the global Rademacher complexity summarizes the singular values of the local model weight W_k for shallow linear models, while for deep models it removes the restriction $P(\ell_f - \ell_{f^*})^2 \leq r$ from local Rademacher complexity in (2). Thus, it is reasonable to compare the generalization performance of these two regularization terms from the local and global Rademacher complexities, respectively.

In TABLE V, we show the results of FedALRC with weighted local Rademacher complexity regularization (FedALRC-l) and weighted Rademacher complexity regularization (FedALRC-g). We observe that FedALRC with weighted local Rademacher complexity regularization performs better, which indicates that our local Rademacher complexity based algorithm has better effect on reducing the excess risk than global Rademacher complexity based methods. This also shows that our generalization theory is practical.

D. Ablation Study

To analyze the influence of weighted local Rademacher complexity and the modified server-side learning rate, we conduct an ablation experiment on non-IID *MNIST* dataset.

In Fig. 2, we compare the performance of FedALRC's counterparts in 100 communication rounds. Compared to FedNova ($\gamma = 1, \alpha = 0$), when $\alpha \neq 0$, we get a higher test accuracy by simply constraining local Rademacher complexity ($\gamma = 1, \alpha = 0.1$), which conincides with our theory that the generalization performance can be improved by reducing the excess risk. By modifying the server-side learning rate ($\gamma = 1.5, \alpha = 0$), the algorithm also performs better than FedNova, where the test accuracy matches the algorithm only with local Rademacher complexity but converges faster. This conincides with the acceleration theory [41] related to the server-side learning rate. FedALRC, not surprisingly, performs the best in both convergence rate and test accuracy. Therefore, we can conclude that a better generalization performance can be obtained by reducing the excess risk, and we can further accelerate the convergence by modifying the server-side learning rate. Furthermore, we can observe that FedALRC seems to be sensitive to the server-side learning rate again. Fortunately, the sensitivity of FedALRC does not impair the performances and we will study the underlying properties in the future work.

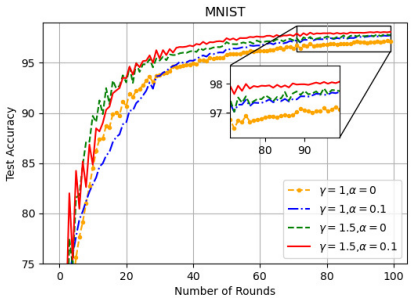

Fig. 2. Results of Ablation Experiments on *MNIST* under Non-IID Setting.

VI. Conclusions and Future Work

In this paper, we present a novel generalization analysis for FL with non-IID data partitioning and derive a sharper risk bound based on weighted local Rademacher complexity. Our theoretical results improve the existing generalization bounds for federated learning, which converges much faster. Motivated by the theoretical findings, we devise effective algorithms to lower the excess risk by constraining local Rademacher complexity, which leads to significant improvements in practice through extensive experiments. Specifically, the convergence rates of the existing risk bounds for horizontal FL methods are usually $\mathcal{O}(1/\sqrt{n})$, while this work achieves a faster rate $\mathcal{O}(1/n)$ by restricting the hypothesis space around the target function. This implies that the proposed learning algorithms can lead to smaller losses with the same number of samples, which is essential to the FL scenarios. The presented proof techniques and algorithmic techniques to control the capacity of the hypothesis are flexible that pave the way to study sharper generalization properties of other FL methods, including vertical FL, domain adaption, multi objectives learning [43] and so on. In the future, we hope to study the underlying properties of the sensitivity of `FedALRC`, and transfer the theoretical findings and improved algorithms in this paper to vertical FL and domain adaption.

Acknowledgments

This work was supported in part by Excellent Talents Program of Institute of Information Engineering, CAS, Special Research Assistant Project of CAS (No. E0YY231114), Beijing Outstanding Young Scientist Program (No.BJJWZYJH01 2019100020098), National Natural Science Foundation of China (No.62076234, No.62106257) and Beijing Municipal Science and Technology Commission under Grant Z191100007119002.

Appendix

We present useful propositions for weighted local Rademacher complexity and detailed proofs of the theorems.

Propositions for Local Rademacher Complexity

***Proposition** 1:* Let ψ be a sub-root function and r^* be the fixed point of ψ, we assume that for any $r > r^*$, ψ satisfies $\psi(r) \geq \sqrt{2}LB\mathcal{R}(\mathcal{H}_r, p) \geq B\mathcal{R}(\mathcal{L}_r, p)$. We define $\mathcal{L}'$ as follows:

$$\mathcal{L}' = \{\ell'_f | \ell'_f = \frac{r\ell_f}{\max(r, P\ell_f^2)}, \ell_f \in \mathcal{L}\}. \tag{3}$$

Then, we have $\mathcal{L}' \subseteq \mathcal{L}_r$.

Proof 1 (Proof of Proposition 1): We consider the following two cases: 1) $P\ell_f^2 \leq r$ and 2) $P\ell_f^2 > r$.

In 1), we have $\ell'_f = \ell_f$ related to (3), so it holds that $P\ell'^2_f = P\ell_f^2 \leq r$.

In 2), we have $\ell'_f = \frac{r\ell_f}{P\ell_f^2}$, so the following inequality holds:

$$P\ell'^2_f = \frac{r^2}{(P\ell_f^2)^2} P\ell_f^2 \leq \frac{r}{P\ell_f^2} P\ell_f^2 = r.$$

The proof is completed.

***Proposition** 2:* Suppose that $\forall G > 1$, we have

$$\Psi(\mathcal{L}') = \sup_{\ell'_f \in \mathcal{L}'} \left[P\ell'_f - P_n\ell'_f\right] \leq \frac{r}{BG}.$$

Then, $\forall f \in \mathcal{H}$, we have

$$P\ell_f \leq \max\{\frac{G}{G-1} P_n\ell_f, P_n\ell_f + \frac{r}{BG}\}.$$

Proof 2 (Proof of Proposition 2): We consider the following two cases:

1) $P\ell_f^2 \leq r$ and 2) $P\ell_f^2 > r$.

In 1), we have $\ell'_f = \ell_f$, so it holds that

$$P\ell_f = P\ell'_f \leq P_n\ell'_f + \Psi(\mathcal{L}') = P_n\ell_f + \frac{r}{BG}.$$

In 2), we have $\ell'_f = \frac{r\ell_f}{P\ell_f^2}$, so the following inequalities hold:

$$P\ell_f - P_n\ell_f \leq \Psi(\mathcal{L}) = \frac{P\ell_f^2}{r}\Psi(\mathcal{L}') \leq \frac{BP\ell_f}{r}\frac{r}{BG} = \frac{P\ell_f}{G}.$$

The proof is completed.

Proof of Theorem 1

In the following Lemma, we derive a sharper generalization bound for non-IID FL based on Talagrand's inequality.

***Lemma** 1:* For any $\delta \in (0,1)$, $\forall f \in \mathcal{H}_r$ and $\forall G > 1$, with probability at least $1-\delta$, the following bound holds:

$$P\ell_f \leq \max\{\frac{G}{G-1} P_n\ell_f, P_n\ell_f + a_1 r^* + \frac{a_2}{n}\}, \tag{4}$$

where $a_1 = 800G/B$ and $a_2 = (16BG + 12B)\log(1/\delta)$.

Proof 3 (Proof of Lemma 1): Let $Z = (\mathbf{x}_i^k, y_i^k)_{i=1\ldots n_k}^{k=1\ldots K}$ be the training samples. We define

$$V = \sup_{\ell_f \in \mathcal{L}} \left[\sum_{k=1}^{K} \frac{p_k}{n_k} \sum_{i=1}^{n_k} \left(\mathbb{E}_Z[\ell_f(\mathbf{x}_i^k, y_i^k)] - \ell_f(\mathbf{x}_i^k, y_i^k)\right)\right],$$

and V' by replacing $(\mathbf{x}_j^s, y_j^s)$ in V with $(\mathbf{x'}_j^s, y'_j^s)$. Then,

$$\begin{aligned} V' = \sup_{\ell_f \in \mathcal{L}} \Bigg[& \frac{p_s}{n_s}\left(\mathbb{E}_{Z'}[\ell_f(\mathbf{x'}_j^s, y'_j^s)] - \ell_f(\mathbf{x'}_j^s, y'_j^s)\right) \\ & - \frac{p_s}{n_s}\left(\mathbb{E}_Z[\ell_f(\mathbf{x}_j^s, y_j^s)] - \ell_f(\mathbf{x}_j^s, y_j^s)\right) \\ & + \sum_{k=1}^{K} \frac{p_k}{n_k} \sum_{i=1}^{n_k} \left(\mathbb{E}_Z[\ell_f(\mathbf{x}_i^k, y_i^k)] - \ell_f(\mathbf{x}_i^k, y_i^k)\right)\Bigg], \end{aligned}$$

We introduce $\bar{f}$ as the labeling function that makes Z get the supremum, and it can be shown that, $\forall j, s$,

$$V - V' \leq \frac{p_s}{n_s}\left(\mathbb{E}_Z[\ell_{\bar{f}}(\mathbf{x}_j^s, y_j^s)] - \ell_{\bar{f}}(\mathbf{x}_j^s, y_j^s)\right) - \frac{p_s}{n_s}\left(\mathbb{E}_{Z'}[\ell_{\bar{f}}(\mathbf{x'}_j^s, y'_j^s)] - \ell_{\bar{f}}(\mathbf{x'}_j^s, y'_j^s)\right).$$

Moreover, we define $(V - V')_+ = \max\{V - V', 0\}$, so

$$(V - V')_+^2 \leq \frac{p_s^2}{n_s^2}\left[\left(\mathbb{E}_Z[\ell_{\bar{f}}(\mathbf{x}_j^s, y_j^s)] - \ell_{\bar{f}}(\mathbf{x}_j^s, y_j^s)\right) - \left(\mathbb{E}_{Z'}[\ell_{\bar{f}}(\mathbf{x'}_j^s, y'_j^s)] - \ell_{\bar{f}}(\mathbf{x'}_j^s, y'_j^s)\right)\right]^2.$$

Combined with $\mathbb{E}_{Z'}[\mathbb{E}_{Z'}[\ell_{\bar{f}}(\mathbf{x'}_j^s, y'_j^s)] - \ell_{\bar{f}}(\mathbf{x'}_j^s, y'_j^s)] = 0$, it holds that

$$\begin{aligned}
&\sum_{s=1}^{K}\sum_{j=1}^{n_s}\mathbb{E}_{Z'}[(V - V')_+^2] \\
\leq&\sum_{s=1}^{K}\sum_{j=1}^{n_s}\frac{p_s^2}{n_s^2}\mathbb{E}_{Z'}\left[\left[\left(\mathbb{E}_Z[\ell_{\bar{f}}(\mathbf{x}_j^s, y_j^s)] - \ell_{\bar{f}}(\mathbf{x}_j^s, y_j^s)\right)\right.\right. \\
&\qquad\left.\left.-\left(\mathbb{E}_{Z'}[\ell_{\bar{f}}(\mathbf{x'}_j^s, y'_j^s)] - \ell_{\bar{f}}(\mathbf{x'}_j^s, y'_j^s)\right)\right]^2\right] \\
\leq&\sum_{s=1}^{K}\frac{p_s^2}{n_s^2}\sum_{j=1}^{n_s}\left(\mathbb{E}_Z[\ell_{\bar{f}}(\mathbf{x}_j^s, y_j^s)] - \ell_{\bar{f}}(\mathbf{x}_j^s, y_j^s)\right)^2 \\
&+\sum_{s=1}^{K}\frac{p_s^2}{n_s^2}\sum_{j=1}^{n_s}\mathbb{E}_{Z'}\left[\left(\mathbb{E}_{Z'}[\ell_{\bar{f}}(\mathbf{x'}_j^s, y'_j^s)] - \ell_{\bar{f}}(\mathbf{x'}_j^s, y'_j^s)\right)^2\right] \\
\leq&\underbrace{\sup_{\ell_f\in\mathcal{L}}\left[\sum_{s=1}^{K}\frac{p_s^2}{n_s^2}\sum_{j=1}^{n_s}\left(\mathbb{E}_Z[\ell_f(\mathbf{x}_j^s, y_j^s)] - \ell_f(\mathbf{x}_j^s, y_j^s)\right)^2\right]}_{\mathbf{V}_1} \\
&+\underbrace{\sup_{\ell_f\in\mathcal{L}}\left[\sum_{s=1}^{K}\frac{p_s^2}{n_s^2}\sum_{j=1}^{n_s}\mathbb{E}_Z\left[\left(\mathbb{E}_Z[\ell_f(\mathbf{x}_j^s, y_j^s)] - \ell_f(\mathbf{x}_j^s, y_j^s)\right)^2\right]\right]}_{\mathbf{V}_2}.
\end{aligned}$$

According to [38], $\forall\xi \in (0, \frac{1}{\theta})$ $(\theta > 0)$, we have the following inequality:

$$\log\mathbb{E}\left[e^{\xi(V-\mathbb{E}[V])}\right] \leq \frac{\xi\theta}{1-\xi\theta}\log\mathbb{E}\left[e^{\frac{\xi}{\theta}(\mathbf{V}_1+\mathbf{V}_2)}\right]. \quad (5)$$

We define $\mathbf{V}_1'$ as

$$\mathbf{V}_1' = \sup_{\ell_f\in\mathcal{L}}\left[\sum_{k=1}^{K}\frac{p_k^2}{n_k^2}\sum_{i=1}^{n_k}\left(\mathbb{E}_Z[\ell_f(\mathbf{x}_i^k, y_i^k)] - \ell_f(\mathbf{x}_i^k, y_i^k)\right)^2 - \frac{p_s^2}{n_s^2}\left(\mathbb{E}_Z[\ell_f(\mathbf{x}_j^s, y_j^s)] - \ell_f(\mathbf{x}_j^s, y_j^s)\right)^2\right].$$

Let $\widetilde{f}$ be the labeling function achieving the supremum of $\mathbf{V}_1$, we have

$$\mathbf{V}_1 - \mathbf{V}_1' \leq \frac{p_s^2}{n_s^2}\left(\mathbb{E}_Z[\ell_{\widetilde{f}}(\mathbf{x}_j^s, y_j^s)] - \ell_{\widetilde{f}}(\mathbf{x}_j^s, y_j^s)\right)^2 \leq \frac{B^2}{n^2} = b^2.$$

Similarly, we introduce $\widetilde{f}'$ as the labeling function achieving the supremum of $\mathbf{V}_1'$, we have

$$\mathbf{V}_1 - \mathbf{V}_1' \geq \frac{p_s^2}{n_s^2}\left(\mathbb{E}_Z[\ell_{\widetilde{f}'}(\mathbf{x}_j^s, y_j^s)] - \ell_{\widetilde{f}'}(\mathbf{x}_j^s, y_j^s)\right)^2 \geq 0.$$

Moreover, it can be shown that

$$\begin{aligned}
&\sum_{s=1}^{K}\sum_{j=1}^{n_s}(\mathbf{V}_1 - \mathbf{V}_1') \\
\leq&\sum_{s=1}^{K}\frac{p_s^2}{n_s^2}\sum_{j=1}^{n_s}\left(\mathbb{E}_Z[\ell_{\widetilde{f}}(\mathbf{x}_j^s, y_j^s)] - \ell_{\widetilde{f}}(\mathbf{x}_j^s, y_j^s)\right)^2 \\
=&\sup_{\ell_f\in\mathcal{L}}\left[\sum_{s=1}^{K}\frac{p_s^2}{n_s^2}\sum_{j=1}^{n_s}\left(\mathbb{E}_Z[\ell_f(\mathbf{x}_j^s, y_j^s)] - \ell_f(\mathbf{x}_j^s, y_j^s)\right)^2\right] = \mathbf{V}_1.
\end{aligned}$$

Therefore, $\frac{\mathbf{V}_1}{b}$ is a b-self bounding function. According to [38], the following inequality holds $\forall\xi \in (0, \frac{1}{b})$:

$$\log\mathbb{E}\left[e^{\xi\frac{\mathbf{V}_1}{b}}\right] \leq \frac{e^{\xi b-1}}{b^2}\mathbb{E}[\mathbf{V}_1] \leq \frac{\xi}{b(1-\xi b)}\mathbb{E}[\mathbf{V}_1]. \quad (6)$$

We further bound $\mathbb{E}[\mathbf{V}_1] - \mathbf{V}_2$ as follows:

$$\begin{aligned}
&\mathbb{E}[\mathbf{V}_1] - \mathbf{V}_2 \\
\leq&\mathbb{E}\left[\sup_{\ell_f\in\mathcal{L}}\left[\sum_{k=1}^{K}\frac{p_k^2}{n_k^2}\sum_{i=1}^{n_k}\left(\mathbb{E}_Z[\ell_f(\mathbf{x}_i^k, y_i^k)] - \ell_f(\mathbf{x}_i^k, y_i^k)\right)^2\right.\right. \\
&\left.\left.-\sum_{k=1}^{K}\frac{p_k^2}{n_k^2}\sum_{i=1}^{n_k}\mathbb{E}_Z\left[\left(\mathbb{E}_Z[\ell_f(\mathbf{x}_i^k, y_i^k)] - \ell_f(\mathbf{x}_i^k, y_i^k)\right)^2\right]\right]\right] \\
\leq&2\mathbb{E}\left[\sup_{f\in\mathcal{H}}\sum_{k=1}^{K}\frac{p_k^2}{n_k^2}\sum_{i=1}^{n_k}\epsilon_i^k\left(\mathbb{E}_Z[\ell_f(\mathbf{x}_i^k, y_i^k)] - \ell_f(\mathbf{x}_i^k, y_i^k)\right)^2\right] \\
\leq&4B\mathbb{E}\left[\sup_{f\in\mathcal{H}}\sum_{k=1}^{K}\frac{p_k^2}{n_k^2}\sum_{i=1}^{n_k}\epsilon_i^k\left(\mathbb{E}_Z[\ell_f(\mathbf{x}_i^k, y_i^k)] - \ell_f(\mathbf{x}_i^k, y_i^k)\right)\right] \\
\leq&\frac{8B}{n}\mathcal{R}(\mathcal{L}, p).
\end{aligned}$$

Substituting the above inequality into (6), we have

$$\log\mathbb{E}\left[e^{\xi\frac{\mathbf{V}_1}{b}}\right] \leq \frac{\xi}{b(1-\xi b)}\left[\frac{8B}{n}\mathcal{R}(\mathcal{L}, p) + \mathbf{V}_2\right]. \quad (7)$$

$\mathbf{V}_2$ can be bounded as

$$\begin{aligned}
\mathbf{V}_2 \leq&\sup_{\ell_f\in\mathcal{L}}\left[\sum_{s=1}^{K}\frac{p_s^2}{n_s}\mathbb{E}\left[\left(\mathbb{E}[\ell_f(\mathbf{x}_m^s, y_m^s)] - \ell_f(\mathbf{x}_m^s, y_m^s)\right)^2\right]\right] \\
\leq&\sup_{\ell_f\in\mathcal{L}}\left[\sum_{s=1}^{K}\frac{p_s^2}{n_s}\mathbb{E}\left[\left(\ell_f(\mathbf{x}_m^s, y_m^s)\right)^2\right]\right] \leq \sum_{s=1}^{K}\frac{p_s^2}{n_s}r \leq \frac{r}{n},
\end{aligned}$$

where $\ell_f(\mathbf{x}_m^s, y_m^s)$ denotes the maximum ℓ_f among all clients.

Combined with the above inequality with (5) and (6), $\forall\xi \in (0, \frac{1}{2b})$, we have

$$\begin{aligned}
&\log\mathbb{E}\left[e^{\xi(V-\mathbb{E}[V])}\right] \\
\leq&\frac{\xi b}{1-\xi b}\left[\frac{\xi}{b(1-\xi b)}\left[\frac{8B}{n}\mathcal{R}(\mathcal{L}, p) + \mathbf{V}_2\right] + \frac{\xi\mathbf{V}_2}{b}\right] \\
\leq&\frac{\xi b}{1-\xi b}\frac{\xi}{b(1-\xi b)}\left[\frac{8B}{n}\mathcal{R}(\mathcal{L}, p) + 2\mathbf{V}_2\right] \\
\leq&\frac{\xi^2}{1-2\xi b}\left[\frac{8B}{n}\mathcal{R}(\mathcal{L}, p) + \frac{2r}{n}\right].
\end{aligned}$$

According to [38], the following inequality holds with probability at least $1-\delta$ ($\delta \in (0,1)$):

$$\begin{aligned} V - \mathbb{E}[V] &\leq \sqrt{4\left[\frac{8B}{n}\mathcal{R}(\mathcal{L},p)+\frac{2r}{n}\right]\log\frac{1}{\delta}} + 2b\log\frac{1}{\delta} \\ &\leq 4\sqrt{\frac{2B\mathcal{R}(\mathcal{L},p)}{n}\log\frac{1}{\delta}} + 2\sqrt{\frac{2r}{n}\log\frac{1}{\delta}} + 2b\log\frac{1}{\delta} \\ &\leq 2\mathcal{R}(\mathcal{L},\mathbf{p}) + \frac{6B}{n}\log\frac{1}{\delta} + 2\sqrt{\frac{2r}{n}\log\frac{1}{\delta}}. \end{aligned}$$

Due to the symmetrization, we have

$$\begin{aligned} &\mathbb{E}[V] \\ =&\mathbb{E}_Z\left[\sup_{\ell_f\in\mathcal{L}}\mathbb{E}_{Z'}\left[\sum_{k=1}^{K}\frac{p_k}{n_k}\sum_{i=1}^{n_k}\left(\ell_f(\mathbf{x'}_i^k,y'^k_i)-\ell_f(\mathbf{x}_i^k,y_i^k)\right)\right]\right] \\ \leq&\mathbb{E}_{Z,Z'}\left[\sup_{\ell_f\in\mathcal{L}}\sum_{k=1}^{K}\frac{p_k}{n_k}\sum_{i=1}^{n_k}\left(\ell_f(\mathbf{x'}_i^k,y'^k_i)-\ell_f(\mathbf{x}_i^k,y_i^k)\right)\right] \\ =&\mathbb{E}_{Z,Z',\epsilon}\left[\sup_{\ell_f\in\mathcal{L}}\sum_{k=1}^{K}\frac{p_k}{n_k}\sum_{i=1}^{n_k}\epsilon_i^k\left(\ell_f(\mathbf{x'}_i^k,y'^k_i)-\ell_f(\mathbf{x}_i^k,y_i^k)\right)\right] \\ \leq&2\mathcal{R}(\mathcal{L},p). \end{aligned}$$

Therefore, with probability at least $1-\delta$ ($\delta \in (0,1)$), the bound for V holds as follows:

$$V \leq 4\mathcal{R}(\mathcal{L},p) + 2\sqrt{\frac{2r}{n}\log\frac{1}{\delta}} + \frac{6B}{n}\log\frac{1}{\delta}. \tag{8}$$

Applying (8) to $\mathcal{L}'$, we have

$$\sup_{\ell'_f\in\mathcal{L}'}\left[P\ell'_f - P_n\ell'_f\right] \leq 4\mathcal{R}(\mathcal{L}',p) + 2\sqrt{\frac{2r}{n}\log\frac{1}{\delta}} + \frac{6B}{n}\log\frac{1}{\delta}. \tag{9}$$

Then, we define τ as the smallest integer satisfies $r\mu^{\tau+1} \geq B^2$ ($\mu > 1$). Combined with the property of Rademacher complexity, we obtain

$$\begin{aligned} &\mathcal{R}(\mathcal{L}',p) = \mathbb{E}\left[\mathbb{E}_\epsilon\left[\sup_{\ell'_f\in\mathcal{L}'}\sum_{k=1}^{K}\frac{p_k}{n_k}\sum_{i=1}^{n_k}\epsilon_i^k\ell'_f(\mathbf{x}_i^k,y_i^k)\right]\right] \\ =&\mathbb{E}\left[\mathbb{E}_\epsilon\left[\sup_{\ell_f\in\mathcal{L}}\sum_{k=1}^{K}\frac{p_k}{n_k}\sum_{i=1}^{n_k}\frac{r}{\max(r,P\ell_f^2)}\epsilon_i^k\ell_f(\mathbf{x}_i^k,y_i^k)\right]\right] \\ \leq&\mathbb{E}\left[\mathbb{E}_\epsilon\left[\sup_{\ell_f\in\mathcal{L}(0,r)}\sum_{k=1}^{K}\frac{p_k}{n_k}\sum_{i=1}^{n_k}\epsilon_i^k\ell_f(\mathbf{x}_i^k,y_i^k)\right]\right] + \\ &\mathbb{E}\left[\mathbb{E}_\epsilon\left[\sup_{\ell_f\in\mathcal{L}(r,B^2)}\sum_{k=1}^{K}\frac{p_k}{n_k}\sum_{i=1}^{n_k}\frac{r}{P\ell_f^2}\epsilon_i^k\ell_f(\mathbf{x}_i^k,y_i^k)\right]\right] \\ \leq&\mathbb{E}\left[\mathbb{E}_\epsilon\left[\sup_{\ell_f\in\mathcal{L}(0,r)}\sum_{k=1}^{K}\frac{p_k}{n_k}\sum_{i=1}^{n_k}\epsilon_i^k\ell_f(\mathbf{x}_i^k,y_i^k)\right]\right] + \\ &\sum_{j=0}^{\tau}\mu^{-j}\mathbb{E}\left[\mathbb{E}_\epsilon\left[\sup_{\ell_f\in\mathcal{L}(r\mu^j,r\mu^{j+1})}\sum_{k=1}^{K}\frac{p_k}{n_k}\sum_{i=1}^{n_k}\epsilon_i^k\ell_f(\mathbf{x}_i^k,y_i^k)\right]\right] \\ \leq&\mathcal{R}(\mathcal{L}_r,p) + \sum_{j=0}^{\tau}\mu^{-j}\mathcal{R}(\mathcal{L}_{r\mu^{j+1}},p) \\ \leq&\frac{\psi(r)}{B} + \frac{1}{B}\sum_{j=0}^{\tau}\mu^{-j}\psi(r\mu^{j+1}). \end{aligned}$$

As a sub-root function, ψ satisfies $\psi(ar) \leq \sqrt{a}\psi(r)$ for any $a > 1$, thus,

$$\mathcal{R}(\mathcal{L}',p) \leq \frac{\psi(r)}{B}\left[1+\sqrt{\mu}\sum_{j=0}^{\tau}\mu^{-\frac{j}{2}}\right] \leq \frac{\psi(r)}{B}\left[1+\frac{\mu}{\sqrt{\mu}-1}\right].$$

Also, we have $\frac{\psi(r)}{\sqrt{r}} \leq \frac{\psi(r^*)}{\sqrt{r^*}}$, so

$$\psi(r) \leq \sqrt{\frac{r}{r^*}}\psi(r^*) = \sqrt{rr^*},$$

By setting $\mu = 4$, then, $\forall r \geq r^*$, we have

$$\mathcal{R}(\mathcal{L}',p) \leq \frac{5\psi(r)}{B} \leq \frac{5\sqrt{rr^*}}{B}. \tag{10}$$

Combined the above inequality with (9), $\forall r \geq r^*$, with probability at least $1-\delta$ ($0 < \delta < 1$), we have

$$\sup_{\ell'_f\in\mathcal{L}'}\left[P\ell'_f - P_n\ell'_f\right] \leq \frac{20\sqrt{rr^*}}{B} + 2\sqrt{\frac{2r}{n}\log\frac{1}{\delta}} + \frac{6B}{n}\log\frac{1}{\delta}. \tag{11}$$

We set $A = \frac{20\sqrt{r^*}}{B} + 2\sqrt{\frac{2}{n}\log\frac{1}{\delta}}$ and $D = \frac{6B}{n}\log\frac{1}{\delta}$, so r is upper bounded by the solution of $A\sqrt{r} + D = \frac{r}{BG}$.

According to [38], we have $r \leq (ABG)^2 + 2BGD$. Thus,

$$\frac{r}{BG} \leq \frac{800Gr^*}{B} + \frac{(16BG+12B)\log(1/\delta)}{n}.$$

This completes the proof.

Based on the definition of $\mathcal{L}^*$, the corresponding localized excess hypothesis space can be defined as

$$\mathcal{H}^* := \{f - f^* | f \in \mathcal{H}\}.$$

Then, we define the weighted local Rademacher complexity of $\mathcal{H}_r^*$ as

$$\mathcal{R}(\mathcal{H}_r^*,p) = \mathcal{R}\{f - f^* | f \in \mathcal{H}, P(\ell_f - \ell_{f^*})^2 \leq r\}.$$

According to [44], if ℓ_f is L-lipschitz for $\mathbb{R}^C$ equipped with the 2-norm, then for any $r > r^*$, it holds that

$$B\mathcal{R}(\mathcal{L}_r^*,p) \leq \sqrt{2}LB\mathcal{R}(\mathcal{H}_r^*,p) \leq \psi(r). \tag{12}$$

Thus, we can apply (4) to $\mathcal{L}_r^*$, which leads to a sharper excess risk bound for FL, and this completes the proof.

Proof of Theorem 2

Let $P\|f-f^*\|_2^2 \leq BP(\ell_f - \ell_{f^*})$, $\forall f \in \mathcal{H}_r^*$, where $B > 1$ is some constant. Thus, we have

$$P(\ell_f - \ell_{f^*})^2 \leq L^2P\|f-f^*\|_2^2 \leq BL^2P(\ell_f - \ell_{f^*}).$$

Due to the convexity of $\mathcal{H}$ and the symmetry of the Rademacher variables, we have

$$\begin{aligned} \mathcal{R}(\mathcal{H}_r^*,p) &= \mathcal{R}\{f-f^*|f\in\mathcal{H}, P(\ell_f-\ell_{f^*})^2 \leq r\} \\ &\leq \mathcal{R}\{f-f^*|f\in\mathcal{H}, P\|f-f^*\|_2^2 \leq \frac{r}{L^2}\} \\ &\leq \mathcal{R}\{f-g|f,g\in\mathcal{H}, P\|f-g\|_2^2 \leq \frac{r}{L^2}\} \\ &= 2\mathcal{R}\{f|f\in\mathcal{H}, P\|f\|_2^2 \leq \frac{r}{4L^2}\} = 2\mathcal{R}(\mathcal{H}'_r,p). \end{aligned}$$

We can rewrite the local weighted Rademacher complexity of $\mathcal{H}_r$ as follows:

$$\begin{aligned}\mathcal{R}(\mathcal{H}'_r,p)&=\mathbb{E}\left[\sup_{f\in\mathcal{H}'_r}\sum_{k=1}^{K}\frac{p_k}{n_k}\sum_{i=1}^{n_k}\sum_{c=1}^{C}\epsilon_{ic}^k f_c(\mathbf{x}_i^k)\right]\\ &=\mathbb{E}\left[\sup_{f\in\mathcal{H}'_r}\sum_{k=1}^{K}\frac{p_k}{n_k}\sum_{i=1}^{n_k}\sum_{c=1}^{C}\epsilon_{ic}^k {\mathbf{W}_{\cdot c}^k}^T\phi(\mathbf{x}_i^k)\right]\\ &=\mathbb{E}\left[\sup_{f\in\mathcal{H}'_r}\sum_{k=1}^{K}p_k\left(\sum_{c=1}^{C}{\mathbf{W}_{\cdot c}^k}^T\frac{1}{n_k}\sum_{i=1}^{n_k}\epsilon_{ic}^k\phi(\mathbf{x}_i^k)\right)\right]\\ &=\mathbb{E}\left[\sup_{f\in\mathcal{H}'_r}\sum_{k=1}^{K}p_k\left\langle\mathbf{W}^k,\mathbf{X}^k\right\rangle\right],\end{aligned}\tag{13}$$

where $\mathbf{W}_{\cdot c}^k$ is the c-th column of $\mathbf{W}^k$, $\mathbf{W}^k,\mathbf{X}^k\in\mathbb{R}^{D\times C}$, and $\mathbf{X}^k$ is defined as $\mathbf{X}^k=[\frac{1}{n_k}\sum_{i=1}^{n_k}\epsilon_{ic}^k\phi(\mathbf{x}_i^k),...,\frac{1}{n_k}\sum_{i=1}^{n_k}\epsilon_{iC}^k\phi(\mathbf{x}_i^k)]$.

Based on the SVD decomposition, we have:

$$\mathbf{W}^k=\sum_j\mathbf{u}_j^k{\mathbf{v}_j^k}^T\lambda_j^k,$$

where $\mathbf{u}_j^k$ and $\mathbf{v}_j^k$ are the column vectors of $\mathbf{U}^k$ and $\mathbf{V}^k$.

Moreover, we have the following inequalities:

$$\begin{aligned}&\langle\mathbf{W}^k,\mathbf{X}^k\rangle\\ &=\left\langle\sum_{j=1}^{\vartheta}\mathbf{u}_j^k{\mathbf{v}_j^k}^T(\lambda_j^k)^2,\sum_{j=1}^{\vartheta}\mathbf{X}^k\mathbf{u}_j^k{\mathbf{u}_j^k}^T(\lambda_j^k)^{-1}\right\rangle\\ &\quad+\left\langle\mathbf{W}^k,\sum_{j>\vartheta}\mathbf{X}^k\mathbf{u}_j^k{\mathbf{u}_j^k}^T\right\rangle\\ &\le\left\|\sum_{j=1}^{\vartheta}\mathbf{u}_j^k{\mathbf{v}_j^k}^T(\lambda_j^k)^2\right\|\left\|\sum_{j=1}^{\vartheta}\mathbf{X}^k\mathbf{u}_j^k{\mathbf{u}_j^k}^T(\lambda_j^k)^{-1}\right\|\\ &\quad+\|\mathbf{W}^k\|\left\|\sum_{j>\vartheta}\mathbf{X}^k\mathbf{u}_j^k{\mathbf{u}_j^k}^T\right\|.\end{aligned}$$

Note that $\mathbb{E}[\phi(\mathbf{x}^k)^T\phi(\mathbf{x}^k)]\le 1$, let

$$\begin{aligned}P\|f\|_2^2&=\mathbb{E}\left[\left\|{\mathbf{W}^k}^T\phi(\mathbf{x}^k)\right\|_2^2\right]=\mathbb{E}\left[\phi(\mathbf{x}^k)^T\mathbf{W}^k{\mathbf{W}^k}^T\phi(\mathbf{x}^k)\right]\\ &\le\mathbb{E}\left[\|\mathbf{W}^k{\mathbf{W}^k}^T\|^2\right]\le\frac{r}{4L^2}.\end{aligned}$$

Thus, we have $\mathbb{E}[\|\mathbf{W}^k{\mathbf{W}^k}^T\|]\le\frac{\sqrt{r}}{2L}$.

Note that $\|\mathbb{E}[\mathbf{W}^k{\mathbf{W}^k}^T]\|\le\mathbb{E}[\|\mathbf{W}^k{\mathbf{W}^k}^T\|]$, according to [37], it holds that:

$$\left\|\sum_{j=1}^{\vartheta}\mathbf{u}_j^k{\mathbf{v}_j^k}^T(\lambda_j^k)^2\right\|\le\|\mathbb{E}[\mathbf{W}^k{\mathbf{W}^k}^T]\|\le\frac{\sqrt{r}}{2L}.\tag{14}$$

Combined with the properties of SVD decomposition, we get

$$\begin{aligned}&\mathbb{E}\left[\left\|\sum_{j=1}^{\vartheta}\mathbf{X}^k\mathbf{u}_j^k{\mathbf{u}_j^k}^T(\lambda_j^k)^{-1}\right\|\right]=\mathbb{E}\left[\sqrt{\sum_{j=1}^{\vartheta}(\lambda_j^k)^{-2}\left\langle\mathbf{X}^k,\mathbf{u}_j^k\right\rangle^2}\right]\\ &\le\sqrt{\sum_{j=1}^{\vartheta}(\lambda_j^k)^{-2}\mathbb{E}\left[\left\langle\mathbf{X}^k,\mathbf{u}_j^k\right\rangle^2\right]}=\sqrt{\frac{\vartheta}{n_k}}.\end{aligned}\tag{15}$$

We also have

$$\mathbb{E}\left[\left\|\sum_{j>\vartheta}\mathbf{X}^k\mathbf{u}_j^k{\mathbf{u}_j^k}^T\right\|\right]\le\sqrt{\frac{\sum_{j>\vartheta}(\lambda_j^k)^2}{n_k}}\le\frac{\sum_{j>\vartheta}\lambda_j^k}{\sqrt{n_k}}.\tag{16}$$

Substituting (14), (15) and (16) into (13), we have

$$\mathcal{R}(\mathcal{H}'_r,p)\le\inf_{\vartheta\ge0}\sum_{k=1}^{K}p_k\left(\frac{1}{2L}\sqrt{\frac{\vartheta r}{n_k}}+\frac{\sum_{j>\vartheta}\lambda_j^k}{\sqrt{n_k}}\right).$$

Thus, the following bound holds:

$$\mathcal{R}(\mathcal{L}_r^*,p)\le\inf_{\vartheta\ge0}\sum_{k=1}^{K}p_k\left(\sqrt{\frac{2\vartheta r}{n_k}}+2\sqrt{2}L\frac{\sum_{j>\vartheta}\lambda_j^k}{\sqrt{n_k}}\right).$$

According to [34], local Rademacher complexity is a sub-root function, so the same argument holds for $\mathcal{R}(\mathcal{L}_r^*,p)$. Therefore, the fixed point r^* of $\mathcal{R}(\mathcal{L}_r^*,\mathbf{p})$ is unique. We set $A=\sum_{k=1}^{K}p_k\sqrt{\frac{2\vartheta}{n_k}}$ and $D=2\sqrt{2}L\sum_{k=1}^{K}p_k\frac{\sum_{j>\vartheta}\lambda_j^k}{\sqrt{n_k}}$, r^* is upper bounded by the solution of $A\sqrt{r}+D=r$, so

$$r^*\le\inf_{\vartheta\ge0}\sum_{k=1}^{K}p_k\left(\frac{2\vartheta}{n_k}+\frac{4\sqrt{2}L\sum_{j>\vartheta}\lambda_j^k}{\sqrt{n_k}}\right),\tag{17}$$

where the last inequality is derived by Jensen's inequality.

Substituting (17) into (1) completes the proof.

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 *Proceedings of International Conference on Artificial Intelligence and Statistics, AISTATS*, vol. 54, 2017, pp. 1273–1282.

[2] H. Hosseini, H. Park, S. Yun, C. Louizos, J. Soriaga, and M. Welling, "Federated learning of user verification models without sharing embeddings," in *Proceedings of International Conference on Machine Learning, ICML*, vol. 139, 2021, pp. 4328–4336.

[3] Z. Yuan, Z. Guo, Y. Xu, Y. Ying, and T. Yang, "Federated deep AUC maximization for hetergeneous data with a constant communication complexity," in *Proceedings of International Conference on Machine Learning, ICML*, vol. 139, 2021, pp. 12 219–12 229.

[4] C. Xu, S. Liu, Z. Yang, Y. Huang, and K.-K. Wong, "Learning rate optimization for federated learning exploiting over-the-air computation," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 12, pp. 3742–3756, 2021.

[5] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated learning on non-iid data silos: An experimental study," in *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. IEEE, 2022, pp. 965–978.

[6] L. Zhang, L. Shen, L. Ding, D. Tao, and L.-Y. Duan, "Fine-tuning global model via data-free knowledge distillation for non-iid federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 174–10 183.

[7] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," in *International Conference on Learning Representations, ICLR*, 2020.

[8] M. R. Glasgow, H. Yuan, and T. Ma, "Sharp bounds for federated averaging (local sgd) and continuous perspective," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 9050–9090.

[9] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in *Advances in Neural Information Processing Systems, NeurIPS*, 2020.

[10] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of Machine Learning and Systems 2020, MLSys*, 2020.

[11] S. J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," in *International Conference on Learning Representations, ICLR*, 2021.

[12] J. Wang, V. Tantia, N. Ballas, and M. G. Rabbat, "Slowmo: Improving communication-efficient distributed SGD with slow momentum," in *International Conference on Learning Representations, ICLR*, 2020.

[13] H. Wang, M. Yurochkin, Y. Sun, D. S. Papailiopoulos, and Y. Khazaeni, "Federated learning with matched averaging," in *International Conference on Learning Representations, ICLR*, 2020.

[14] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "SCAFFOLD: stochastic controlled averaging for federated learning," in *Proceedings of International Conference on Machine Learning, ICML*, vol. 119, 2020, pp. 5132–5143.

[15] K. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust aggregation for federated learning," *IEEE Transactions on Signal Processing*, vol. 70, pp. 1142–1154, 2022.

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

[17] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *Proceedings of International Conference on Machine Learning, ICML*, vol. 139, 2021, pp. 6357–6368.

[18] A. Blum, N. Haghtalab, R. L. Phillips, and H. Shao, "One for one, or all for all: Equilibria and optimality of collaboration in federated learning," in *Proceedings of International Conference on Machine Learning, ICML*, vol. 139, 2021, pp. 1005–1014.

[19] H. Yu, R. Jin, and S. Yang, "On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization," in *Proceedings of International Conference on Machine Learning, ICML*, vol. 97, 2019, pp. 7184–7193.

[20] T. Murata and T. Suzuki, "Bias-variance reduced local SGD for less heterogeneous federated learning," in *Proceedings of International Conference on Machine Learning, ICML*, vol. 139, 2021, pp. 7872–7881.

[21] Z. Li, D. Kovalev, X. Qian, and P. Richtárik, "Acceleration for compressed gradient descent in distributed and federated optimization," in *Proceedings of International Conference on Machine Learning, ICML*, vol. 119, 2020, pp. 5895–5904.

[22] G. Malinovskiy, D. Kovalev, E. Gasanov, L. Condat, and P. Richtárik, "From local SGD to local fixed-point methods for federated learning," in *Proceedings of International Conference on Machine Learning, ICML*, vol. 119, 2020, pp. 6692–6701.

[23] D. Basu, D. Data, C. Karakus, and S. N. Diggavi, "Qsparse-local-sgd: Distributed SGD with quantization, sparsification and local computations," in *Advances in Neural Information Processing Systems, NeurIPS*, 2019, pp. 14 668–14 679.

[24] A. Khaled, K. Mishchenko, and P. Richtárik, "First analysis of local GD on heterogeneous data," *CoRR*, vol. abs/1909.04715, 2019.

[25] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *Proceedings of International Conference on Machine Learning, ICML*, vol. 97, 2019, pp. 4615–4625.

[26] J. Ro, M. Chen, R. Mathews, M. Mohri, and A. T. Suresh, "Communication-efficient agnostic federated averaging," in *Interspeech*, 2021, pp. 871–875.

[27] Y. Mansour, M. Mohri, J. Ro, and A. T. Suresh, "Three approaches for personalization with applications to federated learning," *CoRR*, vol. abs/2002.10619, 2020.

[28] C. Zheng, S. Liu, Y. Huang, and T. Q. Quek, "Privacy-preserving federated reinforcement learning for popularity-assisted edge caching," in *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2021, pp. 01–06.

[29] Y. Wang, Y. Tong, D. Shi, and K. Xu, "An efficient approach for cross-silo federated learning to rank," in *IEEE International Conference on Data Engineering, ICDE*, 2021, pp. 1128–1139.

[30] Z. Zhu, J. Hong, and J. Zhou, "Data-free knowledge distillation for heterogeneous federated learning," in *International Conference on Machine Learning, ICML*, vol. 139, 2021, pp. 12 878–12 889.

[31] X. Ran, L. Ge, and L. Zhong, "Dynamic margin for federated learning with imbalanced data," in *International Joint Conference on Neural Networks, IJCNN*, 2021, pp. 1–8.

[32] B. Wei, J. Li, Y. Liu, and W. Wang, "Federated learning for non-iid data: From theory to algorithm," in *PRICAI 2021*, vol. 13031, 2021, pp. 33–48.

[33] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning," in *Advances in Neural Information Processing Systems, NIPS*, 2007, pp. 161–168.

[34] P. L. Bartlett, O. Bousquet, and S. Mendelson, "Local rademacher complexities," *The Annals of Statistics*, vol. 33, no. 4, pp. 1497–1537, 2005.

[35] Y. Liu and S. Liao, "Eigenvalues ratio for kernel selection of kernel methods," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2015, pp. 2814–2820.

[36] Y. Liu, S. Liao, H. Lin, Y. Yue, and W. Wang, "Infinite kernel learning: Generalization bounds and algorithms," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2017, pp. 2280–2286.

[37] C. Xu, T. Liu, D. Tao, and C. Xu, "Local rademacher complexity for multi-label learning," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1495–1507, 2016.

[38] N. Yousefi, Y. Lei, M. Kloft, M. Mollaghasemi, and G. C. Anagnostopoulos, "Local rademacher complexity-based learning guarantees for multi-task learning," *J. Mach. Learn. Res.*, vol. 19, pp. 38:1–38:47, 2018.

[39] J. Li, Y. Liu, R. Yin, H. Zhang, L. Ding, and W. Wang, "Multi-class learning: From theory to algorithm," in *Advances in Neural Information Processing Systems, NeurIPS*, 2018, pp. 1593–1602.

[40] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[41] H. Yang, M. Fang, and J. Liu, "Achieving linear speedup with partial worker participation in non-iid federated learning," in *International Conference on Learning Representations, ICLR*, 2021.

[42] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[43] C. Cortes, M. Mohri, J. Gonzalvo, and D. Storcheus, "Agnostic learning with multiple objectives," *Advances in Neural Information Processing Systems*, vol. 33, pp. 20 485–20 495, 2020.

[44] C. Cortes, V. Kuznetsov, M. Mohri, and S. Yang, "Structured prediction theory based on factor graph complexity," in *Advances in Neural Information Processing Systems, NIPS*, 2016, pp. 2514–2522.

Bojian Wei received the B.S. degree from China University of Petroleum (East China), Qingdao, China, in 2019. He received the M.S. degree from Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China and School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China, in 2022.

His current research interests include machine learning, federated learning and learning theory.

Jian Li received the Ph.D. degree from Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, in 2020.

He is currently a researcher at the Institute of Information Engineering, Chinese Academy of Sciences. His research interests include statistical machine learning, large-scale kernel methods, and semi-supervised learning.

Yong Liu received the Ph.D. degree in computer science from Tianjin University, Tianjin, China, in 2016.

He is currently an associate researcher at Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China.

Weiping Wang received the Ph.D. degree in computer science from Harbin Institute of Technology, China, in 2008.

He is currently a professor at Institute of Information Engineering, Chinese Academy of Sciences, National Engineering Research Center for Information Security, and the National Engineering Laboratory for Information Security Technology, Beijing, China.