# Federated Learning for Non-IID Data: From Theory to Algorithm

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

Institute of Information Engineering, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Renmin University of China

18th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2021)

November 10, 2021

# Contents

# Non-IID Data Partitioning in Federated Learning

- In Federated Learning (FL) McMahan et al. [2017], the raw data of each client is stored locally and can not be obtained by any third party $\rightarrow$ **Non-IID** problem.
- **Non-IID** (not independently or identically distributed) data partitioning:
  1. The distribution is different on different clients.
  2. The amount of data among different clients is unbalanced.
  3. Data on some clients may be relevant.
- The non-IID problem leads to the decline of the model's effectiveness compared to centralized learning (CL).

# Lack of Generalization Analysis

- Many studies try to solve the non-IID problem by designing new algorithms empirically Wang et al. [2020a], Smith et al. [2017], Pustozerova et al. [2021], McMahan et al. [2017], Li et al. [2020a], Karimireddy et al. [2020], Yu et al. [2019], Wang et al. [2020b], Li et al. [2020c], Briggs et al. [2020], while only a few studies have carried out generalization analysis Mohri et al. [2019] for FL.
  - `FedAvg` applies iterative model averaging to deal with non-IID data.
  - Using local momentum instead of local SGD.
  - Using clustering to FL.
  - Agnostic FL provides a generalization view of FL, but the target is to optimize the worst case in the hypothesis.
- Lack of generalization analysis for FL under the traditional framework.

# Contributions

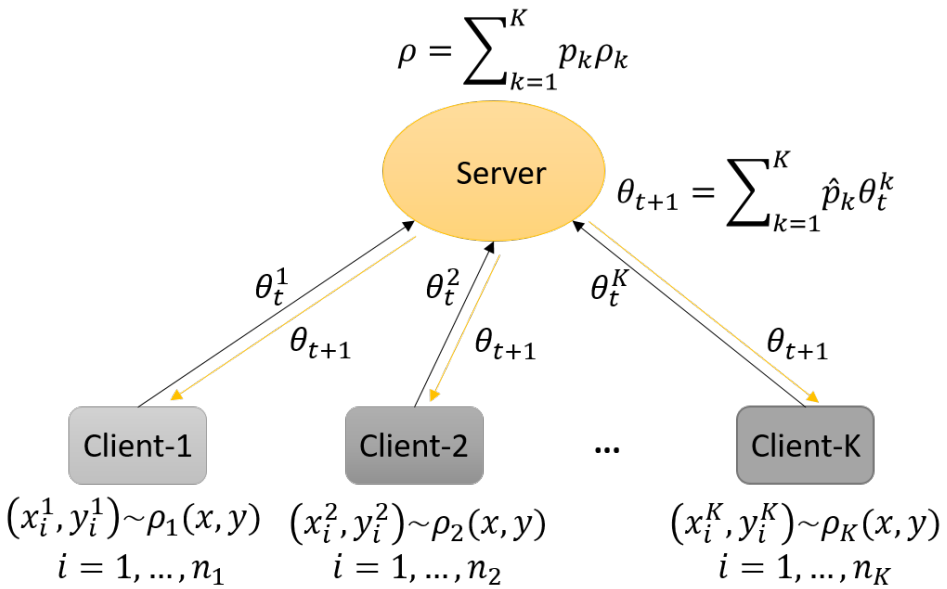
We analyze the ***excess risk*** for FL with non-IID data, which measures the gap between the global model trained by FL and the optimal model trained by CL.

- **Theoretically:** We give the *excess risk* bound between FL on non-IID data and CL for the first time and find out the factors that affect the accuracy decline. We give a reasonable explanation for the bound by decomposing the *excess risk* into three terms: *agnostic error*, *federated error* and *approximation error*.
- **Algorithmically:** We propose a novel algorithm `FedAvgR` (Federated Averaging with Regularization) to improve the performance of FL on non-IID data, which is regularized by Rademacher complexity and discrepancy distance.

# Contents

# Preliminaries and Notations

- Assume that there are $K$ clients in a FL setting, where samples $(\boldsymbol{x}^k, y^k)$ on the $k$-th client with size of $n_k$ are drawn i.i.d. from distribution $\rho_k$, data on different clients may not have the same distribution ($\rho_i \neq \rho_j$), and all clients participate in each communication round.
- The **global distribution** is assumed to be a mixture distribution of local distributions on all $K$ clients: $\rho = \sum_{k=1}^{K} p_k \rho_k$, where $p_k$ is the mixture weight ($\sum_{k=1}^{K} p_k = 1$). Actually, the mixture weight $p_k$ is unknown, so an estimated weight $\widehat{p}_k$ will be applied in practice, which brings us the **estimated global distribution** $\widetilde{\rho} = \sum_{k=1}^{K} \widehat{p}_k \rho_k$.



# Preliminaries and Notations

- The hypothesis space $\mathcal{H} = \{\boldsymbol{x} \to f(\boldsymbol{x})\}$ consists of labeling functions $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ represents the input space and $\mathcal{Y} \subseteq \mathbb{R}^C$ represents the label space. Let $\ell(f(\boldsymbol{x}), y)$ be the loss function, which is assumed to be upper bounded by $M$ ($M > 0$), and $\mathcal{L} = \{\ell(f(\boldsymbol{x}), y) | f \in \mathcal{H}\}$ be the family of loss functions on $\mathcal{H}$, the expected loss of FL on $\rho$ can be described as

$$\mathcal{E}_\rho(f) = \sum_{k=1}^{K} p_k \mathcal{E}_{\rho_k}(f) = \sum_{k=1}^{K} p_k \int_{\mathcal{X}\times\mathcal{Y}} \ell(f(\boldsymbol{x}), y) d\rho_k(\boldsymbol{x}, y),$$

and the corresponding empirical loss is

$$\widehat{\mathcal{E}}_\rho(f) = \sum_{k=1}^{K} p_k \widehat{\mathcal{E}}_{\rho_k}(f) = \sum_{k=1}^{K} p_k \frac{1}{n_k} \sum_{i=1}^{n_k} \ell(f(\boldsymbol{x}_i^k), y_i^k).$$

# Contents

# Error Decomposition

- ***Excess risk***:

$$\mathcal{E}_\rho(\widetilde{f}_{fl}) - \mathcal{E}_\rho(f^*) = \underbrace{\mathcal{E}_\rho(\widetilde{f}_{fl}) - \mathcal{E}_\rho(\widehat{f}_{fl})}_{A_1:=} + \underbrace{\mathcal{E}_\rho(\widehat{f}_{fl}) - \mathcal{E}_\rho(\widehat{f}_{cl})}_{A_2:=} + \underbrace{\mathcal{E}_\rho(\widehat{f}_{cl}) - \mathcal{E}_\rho(f^*)}_{A_3:=}$$

  - $A_1$: *agnostic error*, $A_2$: *federated error*, $A_3$: *approximation error*.
  - $\widetilde{f}_{fl} = \arg\min_{f\in\mathcal{H}} \sum_{k=1}^K \widehat{p}_k \widehat{\mathcal{E}}_{\rho_k}(f)$ is the empirical learner of FL on $\widetilde{\rho}$, $\widehat{f}_{fl} = \arg\min_{f\in\mathcal{H}} \sum_{k=1}^K p_k \widehat{\mathcal{E}}_{\rho_k}(f)$ is the empirical learner on $\rho$, $\widehat{f}_{cl} = \arg\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \ell(f(\boldsymbol{x}_i), y_i)$ is the empirical learner of CL, and $f^* = \arg\min_{f\in\mathcal{H}} \mathcal{E}_\rho(f)$ is the expected (optimal) learner in $\mathcal{H}$ which minimizes the expected loss on $\rho$.
- The labeling function $f$ is formed as $f(\boldsymbol{x}) = \boldsymbol{W}^T\phi(\boldsymbol{x})$,where $\boldsymbol{W} \in \mathbb{R}^{D\times C}$, $\phi(\boldsymbol{x}) \in \mathbb{R}^D$ and $\phi(\cdot)$ is the feature mapping with learnable parameters $\boldsymbol{\varphi}$.

# Bounds of Three Error Terms

To measure the performance gap of a model on different distributed data, we introduce the **discrepancy distance** Mansour et al. [2009] as follows:

$$disc_L(Q_1, Q_2) = \sup_{f \in \mathcal{H}} \left| \mathcal{E}_{Q_1}(f) - \mathcal{E}_{Q_2}(f) \right|,$$

where $Q_1$ and $Q_2$ are two different distributions.

## Theorem (Agnostic Error Bound)

*Assume that $\ell(f(\boldsymbol{x}), y)$ is $\lambda$-$Lipschitz$ equipped with the 2-norm, that is $|\ell(f(\boldsymbol{x}), y) - \ell(f(\boldsymbol{x}'), y')| \leq \lambda \|f(\boldsymbol{x}) - f(\boldsymbol{x}')\|_2$, $B = \sup_{f = \boldsymbol{W}^T \phi(\boldsymbol{x}) \in \mathcal{H}} \|\boldsymbol{W}\|_*$, where $\|\cdot\|_*$ denotes the trace norm. With probability at least $1 - \delta$ ($\delta > 0$):*

$$A_1 \leq 2 disc_L(\widetilde{\rho}, \rho) + 4\sqrt{2}\lambda B \sum_{k=1}^{K} \frac{\widehat{p}_k}{n_k} \sqrt{C} \|\phi(\boldsymbol{X}^k)\|_F + 6M \sqrt{\frac{\mathcal{S}(\widehat{\boldsymbol{p}} \| \bar{\boldsymbol{n}}) \log(2/\delta)}{2n}},$$

*where $\|\phi(\boldsymbol{X}^k)\|_F = \sqrt{\sum_{i=1}^{n_k} \left\langle \phi(\boldsymbol{x}_i^k), \phi(\boldsymbol{x}_i^k) \right\rangle}$, $\mathcal{S}(\widehat{\boldsymbol{p}} \| \bar{\boldsymbol{n}}) = \chi^2(\widehat{\boldsymbol{p}} \| \bar{\boldsymbol{n}}) + 1$, $\chi^2$ denotes the chi-squared divergence, $\widehat{\boldsymbol{p}} = [\widehat{p}_1, ..., \widehat{p}_K]$, and $\bar{\boldsymbol{n}} = \frac{1}{n}[n_1, ..., n_K]$.*

# Proof Sketch

We first decompose $A_1$ into the following parts:

$$A_1 = \mathcal{E}_\rho(\widetilde{f}_{fl}) - \mathcal{E}_{\widetilde{\rho}}(\widetilde{f}_{fl}) + \underbrace{\mathcal{E}_{\widetilde{\rho}}(\widetilde{f}_{fl}) - \mathcal{E}_{\widetilde{\rho}}(\widehat{f}_{fl})}_{A_1' :=} + \mathcal{E}_{\widetilde{\rho}}(\widehat{f}_{fl}) - \mathcal{E}_\rho(\widehat{f}_{fl})$$

We further decompose $A_1'$ as:

$$A_1' = \underbrace{\mathcal{E}_{\widetilde{\rho}}(\widetilde{f}_{fl}) - \widehat{\mathcal{E}}_{\widetilde{\rho}}(\widetilde{f}_{fl})}_{A_{11} :=} + \underbrace{\widehat{\mathcal{E}}_{\widetilde{\rho}}(\widetilde{f}_{fl}) - \widehat{\mathcal{E}}_{\widetilde{\rho}}(\widehat{f}_{fl})}_{A_{12} :=} + \underbrace{\widehat{\mathcal{E}}_{\widetilde{\rho}}(\widehat{f}_{fl}) - \mathcal{E}_{\widetilde{\rho}}(\widehat{f}_{fl})}_{A_{13} :=}$$

- The remain parts of $A_1$ can be bounded by $2disc_L(\rho, \widetilde{\rho})$.
- $\widehat{\mathcal{E}}_{\widetilde{\rho}}(\widetilde{f}_{fl}) \leq \widehat{\mathcal{E}}_\rho(\widehat{f}_{fl}) \quad \Rightarrow \quad A_{12} \leq 0.$

# Proof Sketch

Let $\mathcal{H}$ be a hypothesis space of $f$ defined over $\mathcal{X}$, $\mathcal{L}$ be the family of loss functions associated to $\mathcal{H}$, $\boldsymbol{n} = [n_1, ..., n_K]$ be the vector of sample sizes and $\boldsymbol{p} = [p_1, ..., p_K]$ be the mixture weight vector, the empirical weighted Rademacher complexity of $\mathcal{L}$ is

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

and the empirical weighted Rademacher complexity of $\mathcal{H}$ is

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

where $f_c(\boldsymbol{x}_i^k)$ is the $c$-th value of $f(\boldsymbol{x}_i^k)$ corresponding to the $C$ classes, $\epsilon_i^k$s and $\epsilon_{ic}^k$s are independent Rademacher variables, which are uniformly sampled from $\{-1, +1\}$, respectively.

# Proof Sketch

**$A_{11}$ and $A_{13}$ can be bounded by weighted Rademacher complexity:**

- For any sample $S = \{S_1, ...S_n\}$ drawn from $\rho$, define $\Phi(S)$ by

$$\Phi(S) = \sup_{f \in \mathcal{H}} (\mathcal{E}_{\widetilde{\rho}}(f) - \widehat{\mathcal{E}}_{\widetilde{\rho}}(f)).$$

Let $S' = \{S'_1, ...S'_n\}$ be a sample differing from $S$ only by point ${\boldsymbol{x}'}_i^k$ in $S'_k$ and $\boldsymbol{x}_i^k$ in $S_k$. Then, we have

$$|\Phi(S) - \Phi(S')| \leq \frac{\widehat{p}_k}{n_k} M.$$

Applying McDiarmid's inequality and Jensen's inequality, we get

$$\Phi(S) \leq 2\widehat{\mathcal{R}}(\mathcal{L}, \widehat{\boldsymbol{p}}) + 3M \sqrt{\frac{\chi^2(\widehat{\boldsymbol{p}} || \bar{\boldsymbol{n}}) + 1}{2n} \log \frac{2}{\delta}}.$$

- $A_{11}, A_{13} \leq \Phi(S)$

# Proof Sketch

**Estimate weighted Rademacher complexity:**

- *Lipschitz assumption*: $\widehat{\mathcal{R}}(\mathcal{L}, \widehat{\boldsymbol{p}}) \leq \sqrt{2}\lambda \widehat{\mathcal{R}}(\mathcal{H}, \widehat{\boldsymbol{p}})$.
- Rewriting $\widehat{\mathcal{R}}(\mathcal{H}, \widehat{\boldsymbol{p}})$:

$$\widehat{\mathcal{R}}(\mathcal{H}, \widehat{\boldsymbol{p}}) = \mathbb{E}_{\epsilon}\left[\sup_{f \in \mathcal{H}} \sum_{k=1}^{K} \frac{\widehat{p}_k}{n_k} \langle \boldsymbol{W}_k, \boldsymbol{\Phi}_k \rangle\right],$$

  where $\boldsymbol{W}_k, \boldsymbol{\Phi}_k = [\sum_{i=1}^{n_k} \epsilon_{i1}^k \phi(\boldsymbol{x}_i^k), ..., \sum_{i=1}^{n_k} \epsilon_{iC}^k \phi(\boldsymbol{x}_i^k)] \in \mathbb{R}^{D \times C}$ and $\langle \boldsymbol{W}_k, \boldsymbol{\Phi}_k \rangle = \mathrm{Tr}(\boldsymbol{W}_k^T \boldsymbol{\Phi}_k)$.

- *Hölder's inequality*: $\widehat{\mathcal{R}}(\mathcal{H}, \widehat{\boldsymbol{p}}) \leq B \sum_{k=1}^{K} \frac{\widehat{p}_k}{n_k} \sqrt{\mathbb{E}_{\epsilon}[\|\boldsymbol{\Phi}_k\|_F^2]}$.
- *Jensen's inequality*: $\mathbb{E}_{\epsilon}[\|\boldsymbol{\Phi}_k\|_F^2] \leq C \|\phi(\boldsymbol{X}^k)\|_F^2$.

# Bounds of Three Error Terms

## Theorem (Federated Error Bound)

*Under the same assumptions as the Theorem before, with probability at least* $1-\delta(\delta>0)$*, we have:*

$$A_2 \leq \sum_{k=1}^{K} p_k \left( disc_L(\rho_k, \rho) + \frac{4\sqrt{2}\lambda B}{n_k}\sqrt{C}\|\phi(\boldsymbol{X}^k)\|_F \right) + \sum_{k=1}^{K} p_k \left( 6M\sqrt{\frac{\log(2/\delta)}{2n_k}} \right).$$

## Theorem (Approximation Error Bound)

*Under the same assumptions as the Theorem before, with probability* $1-\delta(\delta>0)$*, we have:*

$$A_3 \leq \frac{4\sqrt{2}\lambda B}{n}\sqrt{C}\|\phi(\boldsymbol{X})\|_F + 3M\sqrt{\frac{\log(2/\delta)}{2n}},$$

*where* $\|\phi(\boldsymbol{X})\|_F = \sqrt{\sum_{i=1}^{n}\langle\phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_i)\rangle}$.

# Proof Sketch

Note that $A_2 = \sum_{k=1}^{K} p_k [\underbrace{\mathcal{E}_{\rho_k}(\widehat{f}_{fl}) - \mathcal{E}_{\rho}(\widehat{f}_{cl})}_{A_2' :=}]$, we decompose $A_2'$ as:

$$\underbrace{\mathcal{E}_{\rho_k}(\widehat{f}_{fl}) - \widehat{\mathcal{E}}_{\rho_k}(\widehat{f}_{fl})}_{A_{21}} + \underbrace{\widehat{\mathcal{E}}_{\rho_k}(\widehat{f}_{fl}) - \widehat{\mathcal{E}}_{\rho_k}(\widehat{f}_{cl})}_{A_{22}} + \underbrace{\widehat{\mathcal{E}}_{\rho_k}(\widehat{f}_{cl}) - \mathcal{E}_{\rho_k}(\widehat{f}_{cl})}_{A_{23}} + \underbrace{\mathcal{E}_{\rho_k}(\widehat{f}_{cl}) - \mathcal{E}_{\rho}(\widehat{f}_{cl})}_{A_{24}}$$

Substituting $A_{22}$ into the equation of $A_2$, due to the definition of $\widehat{f}_{fl}$, we have

$$\sum_{k=1}^{K} p_k [\widehat{\mathcal{E}}_{\rho_k}(\widehat{f}_{fl}) - \widehat{\mathcal{E}}_{\rho_k}(\widehat{f}_{cl})] \leq 0$$

- $A_{21}$ and $A_{23}$ can be bounded by Rademacher complexity.
- $A_{24}$ can be bounded by $disc_L(\rho_k, \rho)$.
- $A_3$ is a constant multiple of the generalization bound for CL, which can be bounded by Rademacher complexity, as well.

# Excess Risk Bound

## Theorem (Excess Risk Bound)

*Under the same assumptions as the Theorem before, With probability at least* $1-\delta$ *(*$\delta > 0$*), the excess risk bound of federated learning on non-IID data holds as follows:*

$$\mathcal{E}_\rho(\widehat{f}_{fl}) - \mathcal{E}_\rho(f^*) \leq \mathcal{O}\left(G_1 + G_2 + G_3\right),$$

*where* $G_1 = disc_L(\widetilde{\rho}, \rho) + \sum_{k=1}^{K} \frac{\widehat{p}_k B\sqrt{C}}{n_k} \|\phi(\boldsymbol{X}^k)\|_F + \sqrt{\frac{\mathcal{S}(\widehat{\boldsymbol{p}}||\bar{\boldsymbol{n}})}{n}}$,

$G_2 = \frac{B\sqrt{C}}{n} \|\phi(\boldsymbol{X})\|_F + \sqrt{\frac{1}{n}}$ *and*

$G_3 = \sum_{k=1}^{K} p_k [disc_L(\rho_k, \rho) + \frac{B\sqrt{C}}{n_k} \|\phi(\boldsymbol{X}^k)\|_F + \sqrt{\frac{1}{n_k}}]$.

# Excess Risk Bound

- According to the bound, we can lower the *excess risk* by reducing $disc_L(\rho_k, \rho)$, $\|\boldsymbol{W}\|_*$, $\|\phi(\boldsymbol{X}^k)\|_F$, and $disc_L(\widetilde{\rho}, \rho)$.
- In non-IID condition, samples on different clients are drawn from different distributions, so the gap between $\rho_k$ and $\rho$ certainly exists. Furthermore, $p_k$ is unknown, **how can we reduce** $disc_L(\widetilde{\rho}, \rho)$**?**

  $$disc_L(\rho_k, \rho) \downarrow \quad \Rightarrow \quad disc_L(\widetilde{\rho}, \rho) \downarrow$$

  Especially, when $\rho_k = \rho$, whatever value we choose for $\widehat{p}_k$, it's not going to make big difference to the global distribution. Therefore, we are able to lower the *excess risk* by reducing $disc_L(\rho_k, \rho)$, $\|\boldsymbol{W}\|_*$ and $\|\phi(\boldsymbol{X}^k)\|_F$.
- **Corollaries** ($\phi(\cdot)$ is upper bounded by $\kappa^2$ and $\widehat{p}_k = p_k$)
  - $disc_L(\rho_k, \rho) = 0$: $\mathcal{O}\left((\kappa B\sqrt{C}+1)\sum_{k=1}^{K}\widehat{p}_k\sqrt{\frac{1}{n_k}}\right)$
  - $n_k = n/K$: $\mathcal{O}\left(\kappa B\sqrt{KC/n}\right)$ (distributed learning)
  - $K = 1$: $\mathcal{O}(\kappa B\sqrt{C/n})$ (centralized learning)

# Contents

# FedAvgR: Federated Averaging with Regularization

1. We choose MMD (Maximum Mean Discrepancy) Borgwardt et al. [2006] to measure the distance between different distributions $Q_1$ and $Q_2$, which is formed as

$$\text{MMD}[Q_1, Q_2] = \sup_{f \in \mathcal{H}} \left( \mathbb{E}_{Q_1}[f(\boldsymbol{x})] - \mathbb{E}_{Q_2}[f(\boldsymbol{x})] \right).$$

   The local distribution $\rho_k$ won't change during training, so we shall reduce the discrepancy after feature mapping. In other words, we can reduce $disc_L(\rho_k^\phi, \rho^\phi)$ instead of $disc_L(\rho_k, \rho)$, where $\rho_k^\phi$ and $\rho^\phi$ are respectively the local feature distribution on client $k$ and global feature distribution.

2. Taking $\text{MMD}[\rho_k^\phi, \rho^\phi]$ as a regularizer with $\|\boldsymbol{W}\|_*$ and $\|\phi(\boldsymbol{X}^k)\|_F$, the objective function on the $k$-th client is

$$\min_{\boldsymbol{W}, \boldsymbol{\varphi}} \frac{1}{n_k} \sum_{i=1}^{n_k} \ell(f(\boldsymbol{x}_i^k), y_i^k) + \alpha\|\boldsymbol{W}\|_* + \beta\|\phi(\boldsymbol{X}^k)\|_F + \gamma\text{MMD}[\rho_k^\phi, \rho^\phi].$$

# FedAvgR: Federated Averaging with Regularization

**Algorithm 1** `FedAvgR`. $K$ clients are indexed by $k$, $\mathcal{B}$ is the local mini-batch size, $E$ is the number of local epochs, $\eta$ is the learning rate, $\boldsymbol{F}$ represents the objective function.

**Server-Aggregate**

initialize $\boldsymbol{W}_0$ and $\boldsymbol{\varphi}_0$
**for** $k = 1, ..., K$ **do**
  $\widehat{\rho}_k^{\phi} \leftarrow$ estimate the distribution of $\phi(\boldsymbol{x})$
  upload the parameters of $\widehat{\rho}_k^{\phi}$ to the server
**end for**
get the global distribution $\widehat{\rho}^{\phi} = \sum_{k=1}^{K} \widehat{p}_k \widehat{\rho}_k^{\phi}$
**for** each round $t = 1, 2, ...$ **do**
  **for** each client $k$ **do**
    $\boldsymbol{W}_{t+1}^k, \boldsymbol{\varphi}_{t+1}^k, \widehat{\rho}_k^{\phi} \leftarrow$ Client-Update$(k, \boldsymbol{W}_t, \boldsymbol{\varphi}_t, \widehat{\rho}^{\phi})$
  **end for**
  update the global distribution $\widehat{\rho}^{\phi}$
  $\boldsymbol{W}_{t+1} \leftarrow \sum_{k=1}^{K} \widehat{p}_k \boldsymbol{W}_{t+1}^k, \boldsymbol{\varphi}_{t+1} \leftarrow \sum_{k=1}^{K} \widehat{p}_k \boldsymbol{\varphi}_{t+1}^k$
**end for**

**Client-Update**$(k, \boldsymbol{W}_t, \boldsymbol{\varphi}_t, \widehat{\rho}^{\phi})$

draw samples $Z_{\widehat{\rho}^{\phi}}$ from $\widehat{\rho}^{\phi}$
**for** epoch$= 1, ..., E$ **do**
  **for** $(\boldsymbol{x}, y) \in \mathcal{B}$ **do**
    calculate MMD$[\widehat{\rho}_k^{\phi}, \widehat{\rho}^{\phi}]$ by $(\boldsymbol{x}, y)$ and $Z_{\widehat{\rho}^{\phi}}$
    $\boldsymbol{F} = \frac{1}{\mathcal{B}} \sum_{(\boldsymbol{x},y)\in\mathcal{B}} \ell(f(\boldsymbol{x}), y) + \alpha\|\boldsymbol{W}\|_* + \beta\|\phi(\boldsymbol{X})\|_F + \gamma \text{MMD}[\widehat{\rho}_k^{\phi}, \widehat{\rho}^{\phi}]$
    $\boldsymbol{W}_{t+1}^k \leftarrow \boldsymbol{W}_t - \eta \nabla_{\boldsymbol{W}_t} \boldsymbol{F}, \boldsymbol{\varphi}_{t+1}^k \leftarrow \boldsymbol{\varphi}_t - \eta \nabla_{\boldsymbol{\varphi}_t} \boldsymbol{F}$
  **end for**
  $\widehat{\rho}_k^{\phi} \leftarrow$ estimate the distribution of $\phi(\boldsymbol{x})$
**end for**

# Contents

# Experimental Setup

- **Environment:** All the experiments are trained on a Linux_x86_64 server (CPU: Intel(R) Xeon(R) Silver 4214 (RAM: 196 GB) / GPU: NVIDIA GeForce RTX-2080ti).
- **Datasets:**
  1. The synthetic dataset in our experiment is generated related to the method in Li et al. [2020b], where the number of samples $n_k$ on client $k$ follows a power law.
  2. We apply the partitioning method related to McMahan et al. [2017] to some LIBSVM Chang and Lin [20] datasets to get non-IID data. We sort each dataset by the label and divide it into $N/N_s$ shards of size $N_s$, where $N$ is the total number of samples, then we assign each client $2$ shards.
- **Model:** We use Random Fourier Feature Rahimi and Recht [2007] to do the feature mapping $\phi(\cdot)$, which is formed as $\sqrt{2}\cos(\boldsymbol{\omega}^T\boldsymbol{x} + b)$, where $\boldsymbol{\omega}$ is sampled from $\mathcal{N}(0, \sigma^2)$, $\sigma$ is related to the corresponding Gaussian kernel, and $b$ is uniformly sampled from $[0, 2\pi]$.

# Experimental Setup

Table 1: Information of Different Datasets

| Dataset | Class | Training Size | Testing Size | Features |
|---|---|---|---|---|
| a1a | 2 | 1605 | 30956 | 123 |
| svmguide1 | 2 | 3089 | 4000 | 4 |
| splice | 2 | 1000 | 2175 | 60 |
| vehicle | 4 | 500 | 446 | 18 |
| dna | 3 | 2000 | 1186 | 180 |
| pendigits | 10 | 7494 | 3498 | 16 |
| satimage | 6 | 4435 | 2000 | 36 |
| usps | 10 | 7291 | 2007 | 256 |
| MNIST | 10 | 60000 | 10000 | $28 \times 28$ |

# Comparative Experiment

We compare `FedAvgR` with `OneShot` Zhang et al. [2015], `FedAvg` McMahan et al. [2017], `FedProx` Li et al. [2020a] and `FL+HC` Briggs et al. [2020].

- `OneShot` aggregates local models when local trainings converge.
- `FedAvg` iteratively averages local models by $n_k/n$.
- `FedProx` adds the last-round's global model to local training as regularization based on `FedAvg`.
- `FL+HC` uses hierarchical clustering to divide clients into several clusters and applies `FedAvg` separately.

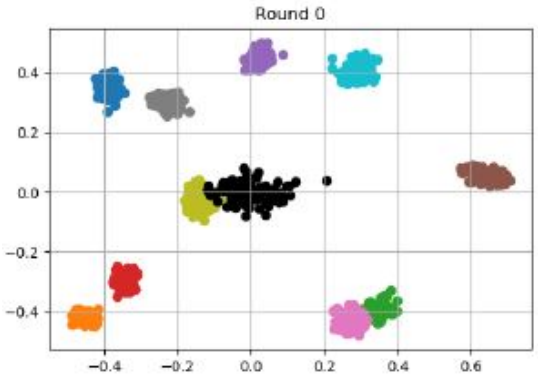
# Comparative Experiment

Table 2: Test Accuracy on Real-World Datasets. We run methods on each dataset 10 times, each with 300 rounds. We bold the numbers of the best method and underline the numbers of other methods which are not significantly worse than the best one.

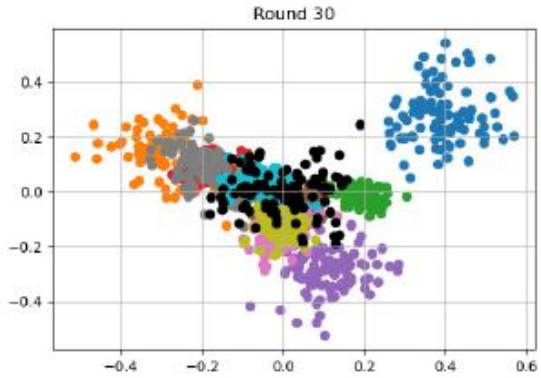
| Dataset | OneShot | FedAvg | FedProx | FL+HC | FedAvgR |
|---|---|---|---|---|---|
| a1a | 76.86±0.30 | <u>84.29±0.06</u> | <u>84.27±0.06</u> | 81.63±0.94 | **84.30±0.06** |
| svmguide1 | 71.50±4.21 | 90.95±0.86 | 91.19±0.84 | 85.66±4.48 | **91.77±1.01** |
| splice | 75.95±4.56 | <u>90.37±0.21</u> | <u>90.38±0.20</u> | 85.12±2.14 | **90.40±0.26** |
| vehicle | 52.31±4.36 | 78.61±1.08 | 78.58±1.06 | 62.24±8.12 | **78.82±0.98** |
| dna | 63.73±1.02 | 95.23±0.17 | 95.18±0.21 | 92.09±3.25 | **95.59±0.23** |
| pendigits | 46.70±2.32 | 94.87±0.58 | 94.85±0.59 | 86.81±4.58 | **95.12±0.48** |
| satimage | 73.07±2.39 | <u>88.83±0.41</u> | 88.46±0.31 | 76.72±2.96 | **88.93±0.39** |
| usps | 56.83±4.06 | 94.57±0.15 | 94.53±0.13 | 88.03±3.62 | **94.80±0.19** |
| MNIST | 68.80±2.06 | 97.26±0.09 | 97.24±0.07 | 85.13±2.23 | **97.34±0.06** |

# Impact of MMD$[\rho_k^\phi, \rho^\phi]$

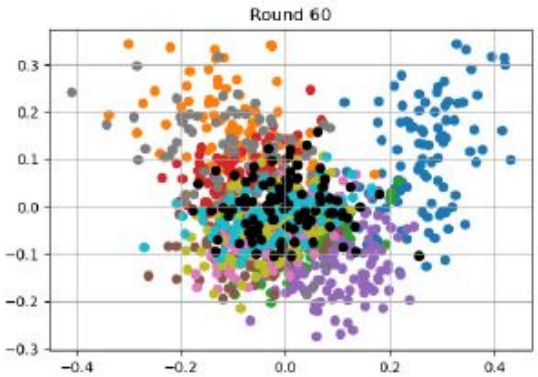
We run 100 rounds on the synthetic dataset with $(u, v) = (1, 1)$ and sample 100 points from each $\widehat{\rho}_k^\phi$ and $\widehat{\rho}^\phi$ to show the impact of MMD$[\rho_k^\phi, \rho^\phi]$.



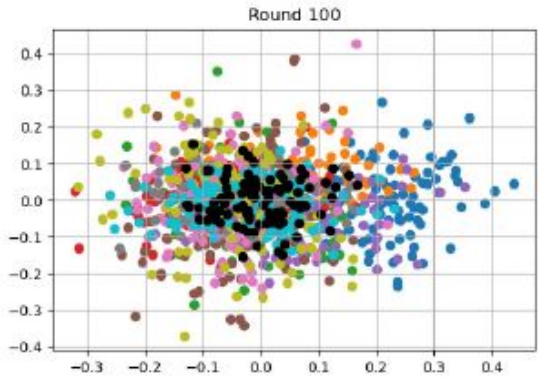

(a) Round 0: 0.766




(b) Round 30: 0.098




(c) Round 60: 0.034




(d) Round 100: 0.019

# Impacts of Different Regularizers

We run 250 rounds on the synthetic dataset with $(u, v) = (0.5, 0.5)$ and some real-world datasets with non-IID partitioning.

Table 3: Test Accuracy of `FedAvgR` with Different Regualrizers

| Dataset | No Regularizer | $\|\boldsymbol{W}\|_*$ | $\|\phi(\boldsymbol{X}^k)\|_F$ | MMD | All Regularizers |
|---|---|---|---|---|---|
| svmguide1 | 89.05 | 89.20 | 89.45 | 89.61 | **90.70** |
| vehicle | 77.12 | 77.17 | 77.17 | 77.46 | **78.32** |
| dna | 95.33 | 95.52 | 95.36 | 95.45 | **95.70** |
| pendigits | 95.70 | 95.71 | 95.74 | **95.94** | 95.90 |
| usps | 94.57 | 94.72 | **94.82** | 94.80 | **94.82** |
| synthetic | 95.82 | 96.07 | 96.06 | 96.12 | **96.23** |

# Contents

# Conclusion

- In this paper, we give an *excess risk* bound for **federated learning on non-IID data** through **Rademacher complexity** and **discrepancy distance**, analyzing the error between it and the optimal centralized learning model. Based on our theory, we propose `FedAvgR` to improve the performance of federated learning in non-IID setting, where three regularizers are added to achieve a sharper bound. Experiments show that our algorithm outperforms the previous methods. As the first work to analyze the *excess risk* under a more general framework, our work will provide a reference for the future study of generalization properties in federated learning with non-IID data. Besides, the proof techniques in this paper are helpful to the research of error analysis related to the distributed framework.

# References


Karsten M. Borgwardt, Arthur Gretton, Malte J. Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alexander J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. In *Proceedings of the 14th International Conference on Intelligent Systems for Molecular Biology*, pages 49–57, 2006.

Christopher Briggs, Zhong Fan, and Peter Andras. Federated learning with hierarchical clustering of local updates to improve training on non-iid data. In *International Joint Conference on Neural Networks, IJCNN*, pages 1–9. IEEE, 2020.

Chih-Chung Chang and Chin-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, pages 2:27:1–27:27, 20.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: stochastic controlled averaging for federated learning. In *ICML*, volume 119, pages 5132–5143, 2020.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *MLSys*, 2020a.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *ICLR*, 2020b.

Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtárik. Acceleration for compressed gradient descent in distributed and federated optimization. In *ICML*, volume 119, pages 5895–5904, 2020c.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT*, 2009.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, volume 54, pages 1273–1282, 2017.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *ICML*, volume 97, pages 4615–4625, 2019.

Anastasia Pustozerova, Andreas Rauber, and Rudolf Mayer. Training effective neural networks on structured data with federated learning. In *AINA*, volume 226, pages 394–406, 2021.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *NIPS*, pages 1177–1184, 2007.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S. Talwalkar. Federated multi-task learning. In *NIPS*, pages 4424–4434, 2017.

Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris S. Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In *ICLR*, 2020a.

Jianyu Wang, Vinayak Tantia, Nicolas Ballas, and Michael G. Rabbat. Slowmo: Improving communication-efficient distributed SGD with slow momentum. In *ICLR*, 2020b.

Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. In *ICML*, volume 97, pages 7184–7193, 2019.

Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.*, 16:3299–3340, 2015.


中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

中国科学院大学
University of Chinese Academy of Sciences

中國人民大學
RENMIN UNIVERSITY OF CHINA

# Thank You