FOCUS: Fairness via Agent-Awareness for Federated Learning on Heterogeneous Data

Wenda Chu[*]
Tsinghua University
chuwd19@mails.tsinghua.edu.cn

Chulin Xie[*]
University of Illinois Urbana-Champaign
chulinx2@illinois.edu

Boxin Wang
University of Illinois Urbana-Champaign
boxinw2@illinois.edu

Linyi Li
University of Illinois Urbana-Champaign
linyi2@illinois.edu

Lang Yin
University of Illinois Urbana-Champaign
langyin2@illinois.edu

Han Zhao
University of Illinois Urbana-Champaign
hanzhao@illinois.edu

Bo Li
University of Illinois Urbana-Champaign
lbo@illinois.edu

Abstract

Federated learning (FL) provides an effective paradigm to train machine learning models over distributed data with privacy protection. However, recent studies show that FL is subject to various security, privacy, and fairness threats due to the potentially malicious and heterogeneous local agents. For instance, it is vulnerable to local adversarial agents who only contribute low-quality data, with the goal of harming the performance of those with high-quality data. This kind of attack hence breaks existing definitions of fairness in FL that mainly focus on a certain notion of performance parity. In this work, we aim to address this limitation and propose a formal definition of *fairness via agent-awareness* for FL (FAA), which takes the heterogeneous data contributions of local agents into account. In addition, we propose a fair FL training algorithm based on agent clustering (FOCUS) to achieve FAA. Theoretically, we prove the convergence and optimality of FOCUS under mild conditions for linear models and general convex loss functions with bounded smoothness. We also prove that FOCUS always achieves higher fairness measured by FAA compared with standard FedAvg protocol under both linear models and general convex loss functions. Empirically, we evaluate FOCUS on four datasets, including synthetic data, images, and texts under different settings, and we show that FOCUS achieves significantly higher fairness based on FAA while maintaining similar or even higher prediction accuracy compared with FedAvg.

1 Introduction

Federated learning (FL) is emerging as a promising approach to enable scalable intelligence over next-generation mobile networks [1, 2]. It transforms the machine learning ecosystem from “centralized over-the-cloud" to “decentralized over-the-edge" in order to (a) alleviate the communication

[*] These authors contributed equally to this work

Preprint. Under review.

bottleneck for pooling massive amounts of data from millions of local users, (b) protect users' privacy by avoiding data egress from their devices, (c) provide personalized intelligent services effectively, and (d) enable large-scale model training.

Despite significant recent milestones in FL, recent studies show that FL is vulnerable to different training-time attacks due to the untrusted local agents [3, 4, 5] and privacy attacks even if there is only one adversarial local agent [6, 7]. Given such untrustworthiness and heterogeneity nature of local agents in FL, especially in the non-IID setting, it is natural to ask: *Can we ensure the fairness of the final learned model for agents?* Indeed, considering the wide application of FL, including medical analysis [8, 9], recommendation systems [10, 11], and personal Internet of Things (IoT) devices [12], it is vitally important to ensure the fairness of FL before its large-scale deployment.

There has been a line of research exploring fairness in federated learning. However, current studies either focus on the fairness of the final trained model regarding the protected attributes without considering different "contributions" of agents [13, 14] or focus on accuracy parity across agents [15, 16, 17]. Several works have taken the contribution of agents into account, while the metric of contribution measurement varies, including the predefined contributions of agents [18], local data quality [18, 19], and local data sizes [20]. In this work, we take into account the heterogeneity of local agents' data and aim to define and enhance the **fairness via agent-awareness for FL (FAA)**. In particular, for FL trained with standard FedAvg [21], if we denote the data of agent e as D_e with n_e the size of D_e and the total number of data as n, the final trained model aims to minimize the loss with respect to the uniform distribution $\mathcal{P} = \sum_{e=1}^{E} \frac{n_e}{n} D_e$, where E is the total number of agents. Obviously, in practice, some local agents may have low-quality data (e.g., free riders), so intuitively it is "unfair" to train the final model regarding such uniform distribution over all agents, which will sacrifice the performance of agents with high-quality data. Thus, in this work, we propose to define the fairness of the local agents in FL via agent-awareness based on the excess risk of each agent e, which stands for the loss of the final model parameterized by θ subtracted by the Bayes error [22] of the local data distribution: $\mathcal{E}_e(\theta) = \mathcal{L}_e(\theta) - \min_w \mathcal{L}_e(w)$. The overall fairness of FL (FAA) is then expressed as: $\mathcal{F}(\theta) = \max_{e_1, e_2 \in E} \left| \mathcal{E}_{e_1}(\theta_{e_1}) - \mathcal{E}_{e_2}(\theta_{e_2}) \right|$, where E denotes the set of local agents.

Based on our definition of fairness in FL via agent awareness (FAA), we propose the *fair FL algorithm based on agent clustering* (FOCUS) to improve the fairness of the trained models. Specifically, we first cluster the local agents based on their data distributions and then train a model for each cluster. During inference time, the final prediction will be the weighted aggregation over the prediction result of each cluster-based model. Theoretically, we prove that the final converged stationary point of FOCUS is exponentially close to the optimal clustering assignment under mild conditions. In addition, we prove that the fairness FAA of FOCUS is strictly higher than that of FedAvg under both linear models and general convex losses. Empirically, we evaluate FOCUS on four datasets, and we show that FOCUS achieves higher fairness measured by FAA than FedAvg, while maintaining similar or even higher prediction accuracy.

Technical contributions. In this work, we focus on defining and improving the fairness of FL by taking the heterogeneous data contributions of local agents into account. We make contributions on both theoretical and empirical fronts.

- We formally define the fairness via agent-awareness for FL (FAA) based on the agent-level excess risks by taking the heterogeneity nature of local data into account.
- We propose a fair FL algorithm based on agent clustering (FOCUS) to improve fairness measured by FAA, especially in the non-IID settings. We prove the convergence rate and optimality of FOCUS under linear models and general convex losses.
- We prove that FOCUS achieves stronger fairness measured by FAA compared with FedAvg on both linear models and general convex losses.
- Empirically, we compare FOCUS with FedAvg on four datasets, including synthetic data, images, and texts under non-IID data settings. We show that FOCUS indeed achieves stronger fairness measured by FAA while maintaining similar or even higher prediction accuracy on all datasets.

2 Related work

Fair Federated Learning Fairness in FL has attracted great attention. Multiple frameworks have attempted to address the fairness issue by enforcing accuracy parity and its variants among agents in

FL. Li et al. [15] first defined agent-level fairness by considering the accuracy equity across agents and achieved the fairness by regularizing the agents with worse performance to have a higher weight in the final objective function. However, this definition of fairness fails to capture the heterogeneous nature of local agents. Mohri et al. [17] pursues accuracy parity by improving the worst-performing agent. In addition, it [23] aims to reduce the disparity during the aggregation step [23], where the server takes care of the gradients with high conflicts (e.g., have negative inner products or magnitudes with large differences) before aggregation in each round. [18] predefines the agent contribution levels based on an a priori assessment of data, which lacks quantitative measurement metrics in practice, while we estimate the agents' inherent distribution directly based on the model performance.

Clustered Federated Learning Clustered FL algorithm is initially designed for multitasking and personalized federated learning, which assumes that agents can be naturally partitioned into clusters [24, 25, 26, 27]. The existing clustering criterion includes the clustering mechanism, which aims to achieve the lowest loss [24], or to optimize the clustering center to be close to the local model [25]. Another common metric is the gradient similarity, where agents with similar gradient updates (with respect to, e.g., cosine similarity [26]) are assigned to the same group. Besides the more straightforward hard clustering, soft clustering has also been proposed [27, 28, 29, 30], which enables the agents to benefit from multiple parties. However, none of these works considers fairness and its implications, and our work will make the first attempt to bridge them.

3 Fair Federated Learning on Heterogeneous Data

In this section, we first define fairness via agent-awareness (FAA) under the setting of federated learning with heterogeneous data, and then introduce our fair federated learning based on agent clustering (FOCUS) algorithm to achieve FAA.

3.1 Fairness via Agent-Awareness under FL with Heterogeneous Data

Given a set of E agents participated in the FL training process, each agent e only has accesses to its local dataset: $D_e = \{(x_e, y_e)\}_{i=1}^{n_e}$, which is sampled from a distribution $\mathcal{P}_e$. The overall goal is to minimize the population loss $\mathcal{L}_E(\theta)$ based on the local loss $\mathcal{L}_e(\theta_e)$ of each agent through communicating privacy-preserving gradient information:

$$\mathcal{L}_E(\theta) = \sum\nolimits_{e \in E} \frac{|D_e|}{n} \mathcal{L}_e(\theta_e), \qquad \mathcal{L}_e(\theta_e) = \mathbb{E}_{(x,y)\in\mathcal{P}_e} \ell(h_{\theta_e}(x), y). \tag{1}$$

where $\ell(\cdot,\cdot)$ is a loss function that measures the difference between model prediction $h_{\theta_e}(x)$ and target label y and n represents the total number of training samples.

In such a training scenario, agents across the federated network may have heterogeneous data, causing the final model to sacrifice the performance for partial or all agents. Regarding this issue, the existing study defines agent-level fairness based on the accuracy equity among agents [15]. However, such a fairness definition fails to capture the heterogeneity of local data distributions. For instance, consider a simple scenario when an agent e samples its data from random noises. The fairness defined by accuracy equity does not provide meaningful measurement since the test accuracy of e cannot be improved anyway. Moreover, adopting the fairness metric of accuracy equity among agents as a training objective might lead to unforeseen subsequences for agents with heterogeneous data. Intuitively, the performance of the agents with high-quality data distribution (e.g., clean or better generality) can be severely compromised by the agents with low-quality data (e.g., noisy or lower generality). This will not only impair the overall performance of the aggregated model but also lead to unfair performance for agents with high-quality data. That is to say, the measurement of fairness in FL should be able to recognize and characterize the distinctions of data distributions (contributions) among agents. To provide such a fairness definition for FL considering the contribution of local agents, we define FAA: fairness via agent-awareness for federated learning as follows:

Definition 1 (**Fairness via agent-awareness for FL (FAA)**). *Suppose a set of agents E takes part in a federated learning framework. The overall fairness among all agents is defined as the maximal excess risk difference between two agents.*

$$\mathcal{F}(\theta) = \max_{e_1, e_2 \in E} \Big| \mathcal{E}_{e_1}(\theta_{e_1}) - \mathcal{E}_{e_2}(\theta_{e_2}) \Big|. \tag{2}$$

where the excess risk for an agent $e \in E$ represents the population loss ($\mathcal{L}_e(\cdot)$) difference between the aggregated model and the Bayes optimal error on the data distribution

$$\mathcal{E}_e(\theta) = \mathcal{L}_e(\theta) - \min_w \mathcal{L}_e(w). \tag{3}$$

Definition 1 is a data-dependent measurement of agent-level fairness. Instead of forcing accuracy equity among all agents regardless of their data distributions, we define agent-level fairness as the equity of excess risks among agents that model the local data contribution. We claim that this definition provides meaningful measurements even in the worst case, i.e., when some agents have random noises as local data, as shown in section 5. We note that **lower FAA indicates stronger fairness among agents** according to the definition.

3.2 Fair Federated Learning on Heterogeneous Data via Clustering (FOCUS)

Method Overview. To achieve the fairness definition FAA defined in Section 3.1, we provide an agent clustering-based FL algorithm (FOCUS) by partitioning agents based on their data properties. The key intuition is that grouping agents with similar data distributions together makes FL fair since it reduces the intra-cluster data heterogeneity. Such principle has also been used for other purposes, such as personalization [27]. We will analyze the fairness achieved by FOCUS and compare it with the standard FedAvg both theoretically (Section 4.2) and empirically (Section 5).

We first elaborate our FOCUS algorithm, which leverages the Expectation-Maximization algorithm for agent clustering. We define M as the number of clusters and E as the number of agents. The goal of FOCUS is to simultaneously optimize the soft clustering labels Π and model weights W. Specifically, $\Pi = \{\pi_{em}\}_{e\in[E],m\in[M]}$ are the dynamic soft clustering labels, representing the estimated probability that agent e belongs to cluster m; $W = \{w_m\}_{m\in[M]}$ represents the model weights for M trained models based on different agent clusters. Suppose there are E agents with datasets $D_1, \dots, D_E$. Our FOCUS algorithm follows a two-step scheme that alternately optimizes Π and W.

E step. Expectation steps update the cluster labels Π given the current estimation of (Π, W). In the k-th communication round, the server broadcasts the M models to all agents and asks them to report the expected training loss $\mathbb{E}_{(x,y)\in D_e}\ell(x, y; w_m^{(k)})$ for each model $m \in [M]$. The server then updates the soft clustering labels Π according to Eq. (8).

M step. The goal of M steps in Eq. (9) is to minimize a weighted sum of empirical losses for all local agents. However, given distributed data, it is impossible to find its exact optimal solution in practice. Thus, we specify a concrete protocol in Eq. (4) $\sim$ Eq. (6) to estimate the objective in Eq. (9). At the k-th M step, the central server broadcasts weights $w_m^{(k)}$ of the M models to every agent. Each agent e first initializes its models $\theta_{em}^{(0)}$ as $w_m^{(k)}$, and then updates the models using its own dataset. To reduce communication costs, each agent is allowed to run SGD locally for T rounds as shown in Eq. (5). After T rounds, each agent should send the updated models $\theta_{em}^{(T)}$ back to the central server; and the server synchronizes the models by a weighted average of all agents. We will provide theoretical analysis for the convergence and optimality of FOCUS considering these multiple local updates in Section 4.

$$\theta_{em}^{(0)} = w_m^{(k)}. \tag{4}$$

$$\theta_{em}^{(t+1)} = \theta_{em}^{(t)} - \eta_t \nabla \sum_{i=1}^{n_e} \ell\Big(h_{\theta_{em}}(x_e^{(i)}), y_e^{(i)}\Big), \forall t = 1, \dots, T-1. \tag{5}$$

$$w_m^{(k+1)} = \sum_{e=1}^{E} \frac{\pi_{em}^{(k+1)} \theta_{em}^{(T)}}{\sum_{e'=1}^{E} \pi_{e'm}^{(k+1)}}. \tag{6}$$

Inference. At inference time, each agent ensembles the M models by a weighted average on their prediction probabilities, i.e., a agent e predicts $\sum_{m=1}^{M} \pi_{em} h_{w_m}(x)$ for input x. Suppose a test dataset D_e^{test} is sampled from distribution $\mathcal{P}_e$. The test loss can be calculated by

$$\mathcal{L}_{test}(W, \Pi) = \frac{1}{|D_e^{test}|} \sum_{(x,y)\in D_e^{test}} \ell\Big(\sum_{m=1}^{M} \pi_{em} h_w(x), y\Big) \tag{7}$$

Algorithm 1 EM clustered federated learning algorithm

```
Input: Data $D_1, \dots, D_E$; $E$ remote agents and $M$ learning models.
  Initialize weights $w_m^{(0)}$ and $\pi_{em}^{(0)} = \frac{1}{M}$ for $m \in [M]$ and $e \in [E]$.
  for $k = 0$ to $K - 1$ do
      for agent $e \in [E]$ do
          for model $m \in [M]$ do
              E step:
                  $$\pi_{em}^{(k+1)} \leftarrow \frac{\pi_{em}^{(k)} \exp\Big(-\mathbb{E}_{(x,y)\in D_e} \ell(x, y; w_m^{(k)})\Big)}{\sum_{m=1}^{M} \pi_{em}^{(k)} \exp\Big(-\mathbb{E}_{(x,y)\in D_e} \ell(x, y; w_m^{(k)})\Big)} \quad (8)$$
          end for
      end for
      for model $m \in [M]$ do
          M step:
              $$w_m^{(k+1)} \leftarrow \arg\min_w \sum_{e=1}^{E} \pi_{em}^{(k+1)} \sum_{i=1}^{n_e} \ell\Big(h_w(x_e^{(i)}), y_e^{(i)}\Big) \quad (9)$$
      end for
  end for
  return model weights $w_m^{(K)}$
```

For unseen agents that do not participate in the training process, their clustering labels Π are unknown. Therefore, an unseen agent e should compute its one-shot clustering label $\pi_{em}^{(1)}, m \in [M]$ according to Eq. (8); and outputs predictions $\sum_{m=1}^{M} \pi_{em}^{(1)} h_{w_m}(x)$ for the test sample x.

4 Theoretical Analysis of FOCUS

In this section, we first present the convergence and optimality guarantees of our EM-based FOCUS algorithm; and then prove it improves the fairness of FL regarding FAA among agents. Our analysis considers linear models with Gaussian data distribution and then extends to nonlinear models with smooth and strongly convex loss functions.

4.1 Convergence Analysis

Linear models We first start with linear models for analysis simplicity. Suppose there are E agents, each with a local dataset $D_e = \{(x_e^{(i)}, y_e^{(i)})\}_{i=1}^{n_e}, (e \in [E])$ generated from a Gaussian distribution. Specifically, we assume each dataset D_e has a mean vector $\mu_e \in \mathbb{R}^d$, so $(x_e^{(i)}, y_e^{(i)})$ is be generated by $y_e^{(i)} = \mu_e^T x_e^{(i)} + \epsilon_e^{(i)}$, where $x_e^{(i)}$ is a random vector $x_e^{(i)} \sim \mathcal{N}(0, \delta^2 I_d)$ and the label $y_e^{(i)}$ is blurred by some random noise $\epsilon_e^{(i)} \sim \mathcal{N}(0, \sigma^2)$.

Each agent is asked to minimize the mean squared error to estimate μ_e, so the empirical loss function for a local agent given dataset D_e is

$$\mathcal{L}_{emp}(D_e; w) = \frac{1}{n_e} \sum_{i=1}^{n_e} (w^T x_e^{(i)} - y_e^{(i)})^2. \quad (10)$$

We further make two assumptions about the heterogeneous agents.

Assumption 1 (Separable distributions)**.** *Suppose there are M predefined vectors $\{c_1, \dots, c_M\}$. These vectors are Separable distributions if for any $m_1, m_2 \in [M]$, $\|c_{m_1} - c_{m_2}\| \geq R$. E agents are divided into M subsets $S_1, \dots, S_M$. For any agent $e \in S_m$, $\|\mu_e - c_m\| \leq r < \frac{R}{2}$.*

Assumption 2 (Proper initialization)**.** *Suppose we train M models and $\pi_{em}^{(0)} = \frac{1}{M}, \forall e, m$. Also assume we pick an initialization w_m for each model $m \in [M]$, such that*

$$\|w_m^{(0)} - c_m\| \leq \alpha = \frac{R}{2} - r - \Delta_0. \quad (11)$$

for some $\Delta_0 > 0$.

Assumption 1 guarantees that the heterogeneous data distributions are separable so that there exists an optimal clustering, in which $\{c_1^*, \dots, c_M^*\}$ are the centers of clusters. We also make assumptions for weights initialization in Assumption 2 to ensure a slight bias for initialized weights $w_m^{(0)}$ towards one of the cluster centers w_m^*.

We present Theorem 1 to demonstrate the linear convergence rate to the optimal cluster centers for FOCUS given Assumption 1 and Assumption 2. Detailed proofs can be found in Appendix A.1.

Theorem 1. *With initialization $\pi_{em}^{(0)} = \frac{1}{M}$ and $\|w_m^{(0)} - c_m\| \leq \frac{R}{2} - r - \Delta_0$ for some $\Delta_0 > 0$, assuming $n_e = O(d)$, if learning rate $\eta \leq \min(\frac{1}{4\delta^2}, \frac{\beta}{T^{3/2}})$, the weights (Π, W) converge by*

$$\pi_{em}^{(k)} \geq \frac{1}{1 + (M-1) \cdot \exp(-2R\delta^2\Delta_0 k)} \tag{12}$$

$$\mathbb{E}\|w_m^{(k)} - c_m\|_2^2 \leq (1 - \frac{2\eta\gamma_m\delta^2}{M})^{kT}(\|w_m^{(0)} - c_m\|_2^2 + A) + \frac{2\eta\gamma_m\delta^2 B}{M - M(1 - \frac{2\eta\gamma_m\delta^2}{M})^T}. \tag{13}$$

where k is the total number of communication rounds, and

$$A = \frac{2ET(M-1)\delta^2}{(1-\frac{2\eta\delta^2\gamma_m}{M})^T - \exp(-2R\delta^2\Delta_0)}, B = 4\eta^{1/3}\beta^{2/3}\delta^2\gamma_m r + 16E\delta^4\beta^2 + \eta^{4/3}\beta^{2/3}\delta^3 EO(\delta^2, \sigma^2). \tag{14}$$

Proof sketch. To prove this theorem, we first consider E steps and M steps separately to derive two lemmas (Lemmas 1 and 2). In E steps, the soft cluster labels π_{em} increase for all $e \in S_m$, as long as $\|w_m^{(k)} - c_m\| < \|w_{m'}^{(k)} - c_m\|, \forall m' \neq m$. On the other hand, $\|w_m^{(k)} - c_m\|$ are guaranteed to shrink linearly as long as π_{em} is large enough for any $e \in S_m$. We then integrate Lemmas 1 and 2 and prove Theorem 1 using an induction statement. □

Remarks. Theorem 1 shows the convergence of parameters (Π, W) to a near optimal solution given linear models. The error term A diminishes exponentially, while the error floor B depends on intra-cluster distribution divergence r and noise level σ. Moreover, this error floor B diminishes as the learning rate η decreases.

Smooth and strongly convex loss functions Next, we extend the convergence analysis to a more general case where the loss functions are L-smooth and μ-strongly convex.

Assumption 3 (Smooth and strongly convex loss functions)**.** *The population loss functions $\mathcal{L}_e(\theta)$ for each agent e is L-smooth,*

$$\|\nabla^2 \mathcal{L}_e(\theta)\|_2 \leq L \tag{15}$$

and μ-strongly convex, i.e., the eigenvalues λ of the Hessian matrix $\nabla^2 \mathcal{L}_e(\theta)$ satisfy:

$$\lambda_{\min}(\nabla^2 \mathcal{L}_e(\theta)) \geq \mu. \tag{16}$$

Assumption 4 (Separable distributions)**.** *E agents are partitioned into M subsets $S_1, \ldots, S_M$. The population loss function $\mathcal{L}_e(\theta)$ of agent e reaches its minimum at θ_e^*. All agents $e \in S_m$ are assumed to share similar data distribution, so that their optimal weights θ_e^* are close to each other:*

$$\|\theta_e^* - w_m^*\| \leq r \tag{17}$$

On the other hand, agents from different subsets have very different data distributions, so

$$\|w_{m_1}^* - w_{m_2}^*\| \geq R, \forall m_1, m_2 \in [M], m_1 \neq m_2. \tag{18}$$

Assumption 5 (Proper Initialization)**.** *Let $\pi_{em}^{(0)} = \frac{1}{M}$ and*

$$\|w_m^{(0)} - w_m^*\| \leq \alpha = \frac{\sqrt{\mu}R}{\sqrt{\mu}+\sqrt{L}} - r - \Delta_0. \tag{19}$$

for some $\Delta_0 > 0$.

Theorem 2. *Suppose loss functions have bounded variance for gradients on local datasets, i.e., $\mathbb{E}_{(x,y)\sim\mathcal{D}_e}[\|\nabla\ell(x, y; \theta) - \nabla\mathcal{L}_e(\theta)\|_2^2] \leq \sigma^2$. Assume population losses are bounded, i.e., $\mathcal{L}_e \in G, \forall e \in [E]$. With initialization $\pi_{em}^{(0)} = \frac{1}{M}$ and $\|w_m^{(0)} - w_m^*\| \leq \frac{\sqrt{\mu}R}{\sqrt{\mu}+\sqrt{L}} - r - \Delta_0$ for some $\Delta_0 > 0$, if each agent chooses learning rate $\eta \leq \min(\frac{1}{2(\mu+L)}, \frac{\beta}{T^{3/2}})$, the weights (Π, W) converges by*

$$\pi_{em}^{(k)} \geq \frac{1}{1 + (M-1)\exp(-\mu R\Delta_0 n)}, \forall t \in S_m \tag{20}$$

$$\mathbb{E}\|w_m^{(k)} - w_m^*\|^2 \leq (1 - \eta A)^{kT}(\|w_m^{(0)} - w_m^*\|^2 + B) + \frac{\eta AC}{1 - (1 - \eta A)^T}. \tag{21}$$

where k is the total number of communication rounds, and

$$A = \frac{2\gamma_m}{M} \frac{\mu L}{\mu + L}, B = \frac{G(M-1)TE(\frac{4L}{\mu} + \frac{6}{\mu(\mu+L)})}{(1-\eta A)^T - \exp(-\mu R \Delta_0)}, \tag{22}$$

$$C = \eta^{1/3}\beta^{2/3}(2\gamma_m Lr\sqrt{\frac{2G}{\mu}} + O(r^2)) + \frac{4EGL^2\beta^2}{\mu} + \eta^{4/3}\beta^{2/3}\frac{E\sigma^2}{n_e}. \tag{23}$$

Proof sketch. We analyze the evolution of parameters (Π, W) for E steps in Lemma 3 and M steps in Lemma 4. Lemma 3 shows that the soft cluster labels π_{em} increase for all $e \in S_m$ in E steps as long as $\|w_m - w_m^*\|_2 < \frac{\sqrt{\mu}R}{\sqrt{\mu}+\sqrt{L}} - r$; whereas Lemma 4 guarantees that the model weights w_m get closer to the optimal solution w_m^* in M steps. We combine Lemmas 3 and 4 together by induction to prove this theorem. Detailed proofs are deferred to Appendix A.2.3. □

Remarks. Theorem 2 extends the convergence guarantee of (Π, W) from linear models (Theorem 1) to general models with smooth and convex loss functions. For any agent e that belongs to a cluster m ($e \in S_m$), its soft cluster label π_{em} converges to 1. Meanwhile, the model weights W converge linearly to the optimal weights of each cluster with an error floor C. Given proper learning rate, the error floor C depends on the intra-cluster distribution deviation r and variance σ, regardless the local iterations in each round T to allow optimal local updates.

4.2 Fairness Analysis

To theoretically show FOCUS achieves higher fairness in FL based on FAA, we focus on a simple yet representative case in which all agents share similar distributions except one outlier agent.

Linear models We first concretize the outlier distribution scenario for linear models. Suppose we have E agents learning weights for M linear models. Their datasets $D_e (e \in [E])$ are generated by $y_e^{(i)} = \mu_e^T x_e^{(i)} - \epsilon_e^{(i)}$ with $x_e^{(i)} \sim \mathcal{N}(0, \delta^2 I_d)$ and $\epsilon_e^{(i)} \sim \mathcal{N}(0, \sigma^2)$. $E-1$ agents learn from normal dataset with ground truth vector $\mu_1, \ldots, \mu_{E-1}$ with $\|\mu_e - \mu^*\|_2 \leq r$, while the E-th agent has an outlier data distribution, with its the ground truth vector μ_E far away from other agents, i.e., $\|\mu_E - \mu^*\|_2 \geq R$.

As we stated in Theorem 1, the model weights (Π, W) converge linearly to the global optimum. Therefore, we analyze the fairness of FOCUS, assuming an optimal (Π, W) is reached.

We compare $\mathcal{F}_{EM}$ with the fairness achieved by the FedAvg algorithm [21] to underscore how agent clustering helps mitigate the unfairness among different agents.

Theorem 3. *When a single agent has outlier distribution, the fairness FAA achieved by FOCUS algorithm with two clusters $M = 2$ is*

$$\mathcal{F}_{EM}(W, \Pi) \leq \delta^2 r^2. \tag{24}$$

while the fairness FAA of FedAvg algorithm is

$$\mathcal{F}_{avg}(W) \geq \delta^2\Big(\frac{R^2(E-2) - 2Rr}{E} + r^2\Big) = \Omega(\delta^2 R^2). \tag{25}$$

Proof sketch. According to Theorem 1, the agents $e \in [E-1]$ with similar distributions converge to the same cluster, producing an aggregated model $w_{m_1} = \frac{\sum_{e=1}^{E} \mu_e}{E-1}$; while the outlier agent is separated from normal agents and train another model w_{m_2} on its own. The detailed proofs are based on these observations and are deferred to Appendix B.1. □

Remarks. When a single outlier exists, the fairness gap between the Fedavg algorithm and FOCUS is shown by Theorem 3.

$$\mathcal{F}_{avg}(W) - \mathcal{F}_{EM}(W, \Pi) \geq \delta^2\Big(\frac{R^2(E-2) - 2Rr}{E}\Big). \tag{26}$$

As long as $R > \frac{2r}{E-2}$, FOCUS is guaranteed to be fairer than Fedavg in terms of FAA. We only discuss the scenario of a single outlier agent here for clarity, but similar conclusions can be drawn for multiple underlying clusters and $M > 2$.

Smooth and strongly convex loss functions We generalize the fairness analysis for linear models to nonlinear models with smooth and convex loss functions. To illustrate the superiority of our FOCUS algorithms in terms of fairness based on FAA, we similarly consider training in the presence of an outlier data distribution.

Suppose we have E agents that learn weights for M models. We assume their population loss functions are L-smooth, μ-strongly convex (as in Assumption 3) and bounded, i.e., $\mathcal{L}_e(\theta) \leq G$. $E-1$ agents learn from similar data distributions. Specifically, we assume the total variation distance between the distributions of any two different agents $i, j \in [E-1]$ is not greater than r: $D_{TV}(\mathcal{P}_i, \mathcal{P}_j) \leq r$. On the other hand, the E-th agent has an outlier data distribution, such that $\mathcal{L}_E(\theta_i^*) - \mathcal{L}_E(\theta_E^*) \geq R$ for any $i \in [E-1]$. We claim that this assumption can be reduced to a lower bound on H-divergence [31] between distributions $\mathcal{P}_i$ and $\mathcal{P}_E$ that $D_H(\mathcal{P}_i, \mathcal{P}_E) \geq \frac{LR}{4\mu}$. (See proofs in Appendix B.3.)

Theorem 4. *The fairness FAA achieved by FOCUS with two clusters $M = 2$ is*

$$\mathcal{F}_{EM}(W, \Pi) \leq \frac{2Gr}{E-1} \tag{27}$$

while the fairness of FedAvg algorithm is

$$\mathcal{F}_{avg}(W) \geq \Big(\frac{E-1}{E} - \frac{L}{\mu E^2}\Big)R - \Big(1 + \frac{L(E-1)}{\mu E} - \frac{L^2}{\mu^2 E}\Big)B - \frac{2L}{\mu E}\sqrt{B(R - \frac{L}{\mu}B)} \tag{28}$$

where $B = \frac{2Gr}{E-1}$.

Proof sketch. According to Theorem 2, agents that normal distributions $e \in [E-1]$ would converge to the same cluster and produce a model $w_{m_1} = \frac{\sum_{e=1}^{E-1} \theta_e^*}{E-1}$.; while the outlier agent trains another model w_{m_2} on its own. We proof Theorem 4 based on these observations in Appendix B.2. □

Remarks. Specifically, when the outlier distribution is very different from the normal distribution, such that $R \gg Gr$ (which means $B \ll R$), we simplify Eq. (28) as

$$\mathcal{F}_{avg}(W) \geq (\frac{E-1}{E} - \frac{L}{\mu E^2})R. \tag{29}$$

Note that $\mathcal{F}_M(W, \Pi) \leq B \ll R$, so the fairness FAA of FedAvg $F_{avg}(W, \Pi)$ is always larger than FAA of FOCUS $\mathcal{F}_M(W)$, as long as $E \geq \sqrt{\frac{L}{\mu}}$.

5 Experimental Evaluation

We conduct extensive experiments on various non-IID data settings to evaluate the fairness measured by FAA for FOCUS and FedAvg [21]. We show that FOCUS achieves significantly higher fairness measured by FAA compared with FedAvg while maintaining similar or even higher accuracy.

5.1 Experimental Setup

Data and Models. We carry out experiments on four different datasets with heterogeneous data settings, ranging from synthetic data for linear models to images (rotated MNIST [32] and rotated CIFAR [33]), to text data for sentiment classification on Yelp [34] and IMDb [35] datasets. We train a fully connected model consisting of two linear layers with ReLU activations for MNIST, a ResNet 18 model [36] for CIFAR, and a pre-trained BERT-base model [37] for the text data. We refer the readers to Appendix C for more implementation details.

Evaluation Metrics. We consider three evaluation metrics: average test accuracy, average test loss, and FAA for fairness. For FedAvg, we evaluate the trained global model on each agent's test data; for FOCUS, we train M models corresponding to M clusters, and use the soft clustering labels $\Pi = \{\pi_{em}\}_{e \in [E], m \in [M]}$ to make aggregated predictions on each agent's test data.

To evaluate FAA of different algorithms, we need to estimate the Bayes optimal loss $\min_w \mathcal{L}_e(w)$ for each local agent e. Thus, we train a centralized model based on each subset of agents with similar data distributions (i.e., the same ground-truth cluster) and use it as a *surrogate* to approximate the Bayes optimum. We select the agents with maximal and minimal excess risks among all, which represents the worst agent pair in terms of fairness, and calculate its gap as FAA (Definition 1). Note that lower FAA indicates stronger fairness by definition.

Table 1: Comparison of FOCUS and FedAvg on different datasets in terms of average test accuracy, average test loss, and fairness FAA. FOCUS achieves stronger fairness measured by FAA compared to FedAvg.

	Synthetic		Rotated MNIST		Rotated CIFAR		Yelp/IMDb	
	FOCUS	FedAvg	FOCUS	FedAvg	FOCUS	FedAvg	FOCUS	FedAvg
Average test accuracy	-	-	**0.953**	0.929	**0.876**	0.843	0.940	0.940
Average test loss	**0.010**	0.031	**0.152**	0.246	**0.272**	0.873	**0.186**	0.236
FAA	**0.001**	0.958	**0.094**	0.363	**0.587**	2.933	**0.064**	0.098

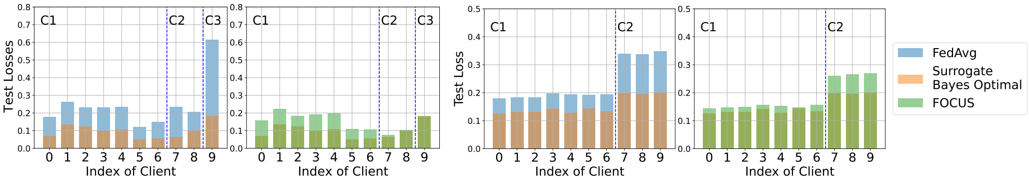

Figure 1: The excess risk of different agents trained with FedAvg and FOCUS. C1, C2 denote cluster 1, cluster2. Left: MNIST; right: sentiment classification on text data.

5.2 Evaluation Results

Synthetic data for linear models. We first evaluate FOCUS on linear regression models with synthetic datasets. We setup $E = 50$ agents with data sampled from Gaussian distributions. Each agent e is assigned with a local dataset of $D_e = \{(x_e^{(i)}, y_e^{(i)})\}_{i=1}^{n_e}$ generated by $y_e^{(i)} = \mu_e^T x_e^{(i)} + \epsilon_e^{(i)}$ with $x_e^{(i)} \sim \mathcal{N}(0, I_d)$ and $\epsilon_e^{(i)} \sim \mathcal{N}(0, \sigma^2)$. We study the case considered in Section 4.2 when a single agent has an outlier data distribution. We set the intra-cluster distance $r = 0.01$ and the inter-cluster distance $R = 1$ in our experiment. Table 1 shows that FOCUS achieves FAA of 0.001, which is much lower than the FAA 0.958 by FedAvg. Note that since it is a regression task, we mainly report the average test loss instead of accuracy here.

Rotated MNIST and CIFAR. Following [24], we rotate the images MNIST and CIFAR datasets with different degrees to create data heterogeneity among agents. Both datasets are evenly split into 10 subsets for 10 agents. For MNIST, 2 subsets are rotated for 90 degrees, 1 subset is rotated for 180 degrees, and the rest 7 subsets are unchanged, yielding an FL setup with three ground-truth clusters. Similarly, for CIFAR, we rotate the images of 3 subsets for 180 degrees, thus creating two clusters.

As shown in Table 1, FOCUS achieves higher average test accuracy, lower average test loss, and lower FAA on both datasets. Fig. 1 (left) shows the surrogate excess risk of every agents on MNIST. We can observe that for the outlier cluster that rotates 180 degrees (i.e., 3rd cluster), the single global model of FedAvg has the highest test loss of 0.61, resulting in high excess risk in the 9th agent. Moreover, the low-quality data of the outlier cluster affect the agents in the 1st cluster, which leads to higher excess risk than that of FOCUS. On the other hand, FOCUS successfully identifies clusters of the outlier distributions, i.e., cluster 2 and 3, rendering models trained from the outlier clusters independent from the normal cluster 1. As shown in Fig. 1, our FOCUS reduces the excess risks of all agents, especially for the outliers. This leads to a more uniform excess risk distribution among agents. Similar trends are also observed in CIFAR, in which our FOCUS reduces the surrogate excess risk for the 9th agent from 2.74 to 0.44. We omit the loss histogram of CIFAR to Appendix C.

Sentiment classification. For the sentiment classification task, Yelp (restaurant reviews) and IMDb (movie reviews) datasets naturally form data heterogeneity among 10 agents and thus create 2 clusters. Specifically, we sample 56k reviews from Yelp datasets distributed among 7 agents and use the whole 25k IMDB datasets distributed among 3 agents to simulate the non-IID data setting.

From Table 1, we can see that while the average test accuracy for FOCUS and FedAvg is close, FOCUS achieves a much lower average test loss. Moreover, our FAA is significantly lower than FedAvg, indicating higher fairness. We also observe that the excess risk on the outlier cluster (i.e., the 2nd cluster) drops significantly than that of FedAvg from Fig. 1 (right).

6 Conclusion

In this work, we provide an agent-level fairness measurement in FL (FAA) by taking the inherent heterogeneous data properties of agents into account. Motivated by our fairness definition in FL, we also provide an effective FL training algorithm FOCUS to achieve high fairness. We theoretically analyze the convergence rate and optimality of FOCUS, and we prove that under mild conditions FOCUS is always more fair than the standard FedAvg protocol. We conduct thorough experiments on synthetic data with linear models as well as image and text datasets on deep neural networks. We show that not only FOCUS achieves stronger fairness than FedAvg, but also FOCUS achieves similar or even higher prediction accuracy across all datasets. We believe our work will inspire new research efforts on exploring the suitable fairness measurements for FL under different requirements.

References

[1] Morgan Ekmefjord, Addi Ait-Mlouk, Sadi Alawadi, Mattias Åkesson, Prashant Singh, Ola Spjuth, Salman Toor, and Andreas Hellander. Scalable federated machine learning with fedn. 2022.

[2] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. 2019.

[3] Arjun Nitin Bhagoji, Supriyo Chakraborty, Seraphin Calo, and Prateek Mittal. Model poisoning attacks in federated learning. In *International Conference on Machine Learning (ICML)*, 2018.

[4] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Analyzing federated learning through an adversarial lens. In *International Conference on Machine Learning (ICML)*, 2019.

[5] Ashwinee Panda, Saeed Mahloujifar, Arjun N. Bhagoji, Supriyo Chakraborty, and Prateek Mittal. Sparsefed: Mitigating model poisoning attacks in federated learning with sparsification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

[6] Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, and Sanjeev Arora. Evaluating gradient inversion attacks and defenses in federated learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[7] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. pages 691–706, 05 2019.

[8] Micah Sheller, Brandon Edwards, G. Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka Colen, and Spyridon Bakas. Federated learning in medicine: Facilitating multi-institutional collaboration without sharing patient data. In *Scientific Reports 10*, 2020.

[9] Mohammed Adnan, Shivam Kalra, Jesse Cresswell, Graham Taylor, and Hamid Tizhoosh. Federated learning and differential privacy for medical image analysis. volume 12, 02 2022.

[10] Lorenzo Minto, Moritz Haller, Hamed Haddadi, and Benjamin Livshits. Stronger privacy for federated collaborative filtering with implicit feedback. In *Fifteenth ACM Conference on Recommender Systems*, pages 342–350, 2021.

[11] Vito Walter Anelli1, Yashar Deldjoo, Antonio Ferrara Tommaso Di Noia, and Fedelucio Narducci. Federated recommender systems with learning to rank. In *29-th Italian Symposium on Advanced Database Systems (SEBD)*, 2021.

[12] Sadi Alawadi, Yuji Dong Victor R. Kebande, Joseph Bugeja, Jan A. Persson, and Carl Magnus Olsson. A federated interactive learning iot-based health monitoring platform. In *European Conference on Advances in Databases and Information Systems*, pages 235–246, 2021.

[13] Lingyang Chu, Lanjun Wang, Yanjie Dong, Jian Pei, Zirui Zhou, and Yong Zhang. Fedfair: Training fair models in cross-silo federated learning. 2021.

[14] Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. Provably fair federated learning via bounded group loss. 2022.

[15] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[16] Kate Donahue and Jon Kleinberg. Models of fairness in federated learning. 2022.

[17] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. 2019.

[18] Jingfeng Zhang, Cheng Li, Antonio Robles-Kelly, and Mohan Kankanhalli. Hierarchically fair federated learning. 2020.

[19] Jiawen Kang, Zehui Xiong, Dusit Niyato, Han Yu, Ying-Chang Liang, and Dong In Kim. Incentive design for efficient federated learning in mobile networks: A contract theory approach. 2019.

[20] Kate Donahue and Jon Kleinberg. Models of fairness in federated learning. 2022.

[21] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. 2017.

[22] Manfred Opper and David Haussler. Generalization performance of bayes optimal classification algorithm for learning a perceptron. *Physical Review Letters*, 66(20):2677, 1991.

[23] Zheng Wang, Xiaoliang Fan, Jianzhong Qi, Chenglu Wen, Cheng Wang, and Rongshan Yu. Federated learning with fair averaging. 2021.

[24] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597, 2020.

[25] Ming Xie, Guodong Long, Tao Shen, Tianyi Zhou, Xianzhi Wang, Jing Jiang, and Chengqi Zhang. Multi-center federated learning. 2021.

[26] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. 2021.

[27] Othmane Marfoq, Giovanni Neglia, Aurélien Bellet, Laetitia Kameni, and Richard Vidal. Federated multi-task learning under a mixture of distributions. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[28] Chengxi Li, Gang Li, and Pramod K. Varshney. Federated learning with soft clustering. 2022.

[29] Yichen Ruan and Carlee Joe-Wong. Fedsoft: Soft clustered federated learning with proximal local updating. 2022.

[30] Morris Stallmann and Anna Wilbik. Towards federated clustering: A federated fuzzy c-means algorithm (ffcm). 2022.

[31] Shengjia Zhao, Abhishek Sinha, Yutong He, Aidan Perreault, Jiaming Song, and Stefano Ermon. Comparing distributions by measuring differences that affect decision making. In *International Conference on Learning Representations*, 2022.

[32] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[33] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[34] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

[35] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[37] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We clearly specify the assumptions that our theoretical results are based on and explain detailed conditions for our experimental evaluations.
   (c) Did you discuss any potential negative social impacts of your work? [Yes] We discuss broader impacts in Appendix D.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See assumptions 1 to 5 in Section 4 for Theorems 1 and 2. The assumptions for Theorems 3 and 4 are clearly listed in Section 4.2.
   (b) Did you include complete proofs of all theoretical results? [Yes] We include complete proofs for all results in Appendices A and B.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We explain our training process in Section 5.1 and discuss more details in Appendix C
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We evaluated various architectures on multiple different datasets where the random seeds are different, and the conclusions are consistent. (See Section 5.)
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix C for details.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] We only use public dataset in our work.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] We only use public dataset in our work.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Contents

A Convergence Proof

A.1 Convergence of Linear Models (Theorem 1)

A.1.1 Key Lemmas

We need to state two lemmas first before proving Theorem 1.

Lemma 1. *Suppose $e \in S_m$ and the m-th cluster is the one closest to c_m. Assume $\|w_m^{(k)} - c_m\| \leq \alpha < \beta \leq \min_{m' \neq m} \|w_{m'}^{(k)} - c_m\|$. Then the E-step updates as*

$$\pi_{em}^{(k+1)} \geq \frac{\pi_{em}^{(k)}}{\pi_{em}^{(k)} + (1 - \pi_{em}^{(k)}) \exp\Big(- (\beta^2 - \alpha^2 - 2(\alpha + \beta) r) \delta^2 \Big)} \tag{30}$$

Remark. Our assumption of proper initialization guarantees that $\|w_m^{(0)} - c_m\| \leq \alpha$ while $\forall m'$, we have $\|w_{m'} - c_m\|_2 \geq \|c_m - \mu_{m'}^*\| - \|w_{m'} - \mu_{m'}^*\| = R - \alpha$. Hence, we substitute $\beta = R - \alpha$ and $\alpha = \frac{R}{2} - r - \Delta$, which yields

$$\pi_{em}^{(k+1)} \geq \frac{\pi_{em}^{(k)}}{\pi_{em}^{(k)} + (1 - \pi_{em}^{(k)}) \exp(-2R\Delta\delta^2)}, \quad \forall e \in S_m \tag{31}$$

For M-steps, the local agents are initialized with $\theta_{em}^{(0)} = w_m^{(k)}$. Then for $t = 1, \ldots, T-1$, each agent use local SGD to update its personal model:

$$\theta_{em}^{(t+1)} = \theta_{em} - \eta_t g_{em}(\theta_{em}) = \theta_{em}^{(t)} - \eta_t \nabla \sum_{i=1}^{n_e} \ell(h_{\theta_{em}}(x_e^{(i)}), y_e^{(i)}). \tag{32}$$

To analyze the aggregated model Eq. (6), we define a sequence of virtual aggregated models $\hat{w}_m^{(t)}$.

$$\hat{w}_m^{(t)} = \sum_{e=1}^{E} \frac{\pi_{em} \theta_{em}^{(t)}}{\sum_{e'=1}^{E} \pi_{e'm}}. \tag{33}$$

Lemma 2. *Suppose any agent $e \in S_m$ has a soft clustering label $\pi_{em}^{(k+1)} \geq p$. Then one step of local SGD updates $\hat{w}_m^{(t)}$ by Eq.* (34)*, if the learning rate $\eta_t \leq \frac{1}{4\delta^2}$.*

$$\mathbb{E}\|\hat{w}_m^{(t+1)} - c_m\|_2^2 \leq (1 - 2\eta_t \gamma_m p \delta^2) \mathbb{E}\|\hat{w}_m^{(t+1)} - c_m\|_2^2 + \eta_t A_1 + \eta_t^2 A_2. \tag{34}$$

$$A_1 = 4\gamma_m r \delta^2 + 2\delta^2 E(1-p), A_2 = 16E(T-1)^2\delta^4 + O(\frac{d}{n_e}) E(\delta^4 + \delta^2 \sigma^2) \tag{35}$$

Remark. Using the recursive relation in Lemma 2, if the learning rate η_t is fixed, the sequence $\hat{w}_m^{(t)}$ has a convergence rate of

$$\mathbb{E}\|\hat{w}_m^{(t)} - c_m\|_2^2 \leq (1 - 2\eta\gamma_m p \delta^2)^t \mathbb{E}\|\hat{w}_m^{(0)} - c_m\|_2^2 + \eta t (A_1 + \eta A_2). \tag{36}$$

A.1.2 Completing the Proof of Theorem 1

We now combine Lemma 1 and Lemma 2 to prove Theorem 1. The theorem is restated below.

Theorem 1. *With initialization $\pi_{em}^{(0)} = \frac{1}{M}$ and $\|w_m^{(0)} - c_m\| \leq \frac{R}{2} - r - \Delta_0$ for some $\Delta_0 > 0$, assuming $n_e = O(d)$, if learning rate $\eta \leq \min(\frac{1}{4\delta^2}, \frac{\beta}{T^{3/2}})$, the weights (Π, W) converge by*

$$\pi_{em}^{(k)} \geq \frac{1}{1 + (M-1) \cdot \exp(-2R\delta^2\Delta_0 k)} \tag{37}$$

$$\mathbb{E}\|w_m^{(k)} - c_m\|_2^2 \leq (1 - \frac{2\eta\gamma_m\delta^2}{M})^{kT} (\|w_m^{(0)} - c_m\|_2^2 + A) + \frac{2\eta\gamma_m\delta^2 B}{M - M(1 - \frac{2\eta\gamma_m\delta^2}{M})^T}. \tag{38}$$

where k is the total number of communication rounds, and

$$A = \frac{2ET(M-1)\delta^2}{(1 - \frac{2\eta\delta^2\gamma_m}{M})^T - \exp(-2R\delta^2\Delta_0)}, B = 4\eta^{1/3}\beta^{2/3}\delta^2\gamma_m r + 16E\delta^4\beta^2 + \eta^{4/3}\beta^{2/3}\delta E O(\delta^4 + \delta^2\sigma^2). \tag{39}$$

Proof. We prove Theorem 1 by induction. Suppose

$$\pi_{em}^{(k)} \geq \frac{1}{1+(M-1)\exp(-2R\delta^2\Delta_0 k)} \tag{40}$$

$$\mathbb{E}\|w_m^{(k)} - c_m\|^2 \leq (1-\frac{2\eta\gamma_m\delta^2}{M})^{kT}(\|w_m^{(0)} - c_m\|^2) + A\Big((1-\frac{2\eta\gamma_m\delta^2}{M})^{kT} - \exp\big(-2R\delta^2\Delta_0 k\big)\Big) + \frac{2\eta\gamma_m\delta^2 B}{M - M(1-\frac{2\eta\gamma_m\delta^2}{M})^T}. \tag{41}$$

□

Then according to Lemma 1,

$$\begin{aligned} \pi_{em}^{(k+1)} &\geq \frac{\pi_{em}^{(k)}}{\pi_{em}^{(k)} + (1-\pi_{em}^{(k)})\exp(-2R\Delta_0\delta^2)} && (42)\\ &\geq \frac{1}{1+(M-1)\exp(-2R\delta^2\Delta_0 n)\exp(-2R\Delta_0\delta^2)} && (43)\\ &\geq \frac{1}{1+(M-1)\exp(-2R\Delta_0\delta^2(k+1))}. && (44) \end{aligned}$$

We recall the virtual sequence of $\hat{w}_m$ defined by Eq. (33). Since models are synchronized after T rounds, the know $\hat{w}_m^{(0)} = w_m^{(k)}$ and $w_m^{(k+1)} = \hat{w}_m^{(T)}$. We then apply Lemma 2 to prove the induction. Note that instead of proving Eq. (38), we prove a stronger induction hypothesis of .

$$\begin{aligned} &\mathbb{E}\|w_m^{(k+1)} - c_m^*\|^2 \\ &= \mathbb{E}\|\hat{w}_m^{(T)} - c_m\|^2 && (45)\\ &\leq (1-2\eta\gamma_m p\delta^2)^T \mathbb{E}\|\hat{w}_m^{(k)} - c_m\|^2 + \eta T(A_1 + \eta A_2) && (46)\\ &\leq (1-2\eta\gamma_m p\delta^2)^T\Big((1-\frac{2\eta\gamma_m\delta^2}{M})^{kT}\|w_m^{(0)} - c_m\|^2 + A((1-\frac{2\eta\gamma_m\delta^2}{M})^{kT} - \exp\big(-2R\Delta_0\delta^2 k\big)) \\ &\qquad + \frac{2\eta\gamma_m\delta^2 B}{M - M(1-\frac{2\eta\gamma_m\delta^2}{M})^T}\Big) + \eta T(4\gamma_m r\delta^2 + 2\delta^2 E(1-p)) + \eta^2 T A_2 && (47)\\ &\leq (1-\frac{2\eta\gamma_m\delta^2}{M})^{(k+1)T}\|w_m^{(0)} - c_m\|^2 \\ &+ \underbrace{A(1-\frac{2\eta\gamma_m\delta^2}{M})^{(k+1)T} - A\exp\big(-2R\Delta_0\delta^2 k\big)(1-\frac{2\eta\gamma_m\delta^2}{M})^T + 2\delta^2 E(1-p)}_{D_1} \\ &+ \underbrace{(1-\frac{2\eta\gamma_m\delta^2}{M})^T\frac{2\eta\gamma_m\delta^2 B}{M - M(1-\frac{2\eta\gamma_m\delta^2}{M})^T} + 4\eta T\gamma_m r\delta^2 + \eta^2 T A_2}_{D_2}. && (48) \end{aligned}$$

Note that $1-p \leq (M-1)\exp\big(-2R\Delta_0\delta^2 k\big)$, so

$$\begin{aligned} D_1 &\leq A(1-\frac{2\eta\gamma_m\delta^2}{M})^{(k+1)T} - A\exp\big(-2R\Delta_0\delta^2 k\big)(1-\frac{2\eta\gamma_m\delta^2}{M})^T + 2\delta^2 ET(M-1)\exp\big(-2R\Delta_0\delta^2 k\big) \\ &\leq A((1-\frac{2\eta\gamma_m\delta^2}{M})^{(k+1)T} - \exp\big(-2R\Delta_0\delta^2(k+1)\big)) && (49) \end{aligned}$$

By $\eta \leq \frac{\beta}{T^{3/2}}$, we have

$$\begin{aligned} D_2 &\leq (1-\frac{2\eta\gamma_m\delta^2}{M})^T\frac{2\eta\gamma_m\delta^2 B}{M - M(1-\frac{2\eta\gamma_m\delta^2}{M})^T} + 4\eta^{1/3}\beta^{2/3}\gamma_m r\delta^2 + 16E\delta^4\beta^2 + \eta^{4/3}\beta^{2/3}O(\delta^4 + \delta^2\sigma^2) \\ &= \frac{2\eta\gamma_m\delta^2 B}{M - M(1-\frac{2\eta\gamma_m\delta^2}{M})^T}. && (50) \end{aligned}$$

Finally we combine Eqs. (48) to (50) so

$$\mathbb{E}\|w_m^{(k+1)-c_m}\|^2 \leq (1-\frac{2\eta\gamma_m\delta^2}{M})^{(k+1)T}\|w_m^{(0)}-c_m\|^2 + A\Big((1-\frac{2\eta\gamma_m\delta^2}{M})^{(k+1)T} - \exp\big(-2R\delta^2\Delta_0(k+1)\big)\Big) + \frac{2\eta\gamma_m\delta^2 B}{M-M(1-\frac{2\eta\gamma_m\delta^2}{M})^T}. \tag{51}$$

Since it is trivial to check that both induction hypotheses hold when $k=0$, Theorem 1 is proved.

A.1.3 Deferred Proofs of Key Lemmas

Lemma 1.

Proof. For simplicity, we abbreviate the model weights $w_m^{(k)}$ by w_m in the proof of this lemma. The n-th E step updates the weights Π by

$$\pi_{em}^{(k+1)} = \frac{\pi_{em}^{(k)}\exp\big[-\mathbb{E}_{(x,y)\sim D_e}({w_m}^Tx-y)^2\big]}{\sum_{m'}\pi_{em'}^{(k)}\exp\big[-\mathbb{E}_{(x,y)\sim D_e}({w_{m'}}^Tx-y)^2\big]} \tag{52}$$

so

$$\begin{aligned}\pi_{em}^{(k+1)} &= \frac{\pi_{em}^{(k)}\exp\big(-\|w_m-\mu_t\|^2\delta^2\big)}{\sum_{m'}\pi_{em'}^{(k)}\exp[-\|w_m'-\mu_t\|^2\delta^2]} &(53)\\ &\geq \frac{\pi_{em}^{(k)}\exp\big(-(\beta-r)^2\delta^2\big)}{\pi_{em}^{(k)}\exp(-(\beta-r)^2\delta^2)+\sum_{m'\neq m}\pi_{em'}^{(k)}\exp(-(\alpha+r)^2\delta^2)} &(54)\\ &\geq \frac{\pi_{em}^{(k)}}{\pi_{em}^{(k)}+(1-\pi_{em}^{(k)})\exp\Big(-(\beta^2-\alpha^2-2(\alpha+\beta)r)\delta^2\Big)} &(55)\end{aligned}$$

□

Lemma 2.

Proof. Notice that local datasets are generated by $X_e \sim \mathcal{N}(0,\delta^2\mathbf{1}^{n_e\times d})$ and $y_e = X_e\mu_e + \epsilon_e$ with $\epsilon_e \sim \mathcal{N}(0,\sigma^2)$. Therefore,

$$\begin{aligned}\|\hat{w}_m^{(t+1)}-c_m\|^2 &= \|w_m^{(t)}-c_m-\eta_t g_t\|^2 &(56)\\ &= \|\hat{w}_m^{(t)}-c_m-\eta_t\frac{2}{n_e}\sum_e\pi_{em}X_e^TX_e(\theta_{em}^{(t)}-\mu_e)+\frac{2\eta_t}{n_e}\sum_e\pi_{em}X_e^T\epsilon_e\|^2 &(57)\\ &= \|\hat{w}_t-c_m-\hat{g}_t\|^2+\eta_t^2\|g_t-\hat{g}_t\|^2+2\eta_t\langle w_t-c_m-\hat{g}_t,\hat{g}_t-g_t\rangle. &(58)\end{aligned}$$

where $\hat{g}_t = \frac{2}{n_e}\sum_e\pi_{em}\mathbb{E}(X_e^TX_e)(\theta_{em}^{(t)}-\mu)$. Since the expectation of the last term in Eq. (58) is zero, we only need to estimate the expectation of $\|\hat{w}_m^{(t)}-c_m-\eta_t\hat{g}_t\|^2$ and $\|\hat{g}_t-g_t\|^2$.

$$\begin{aligned}&\|\hat{w}_m^{(t)}-c_m-\eta_t\hat{g}_t\|^2\\ &= \|\hat{w}_m^{(t)}-c_m\|^2+\frac{4\eta_t^2}{n_e^2}\sum_e\pi_{em}\mathbb{E}(X_e^TX_e)\|\theta_{em}^t-\mu_e\|^2-\frac{4\eta_t}{n_e}\sum_e\pi_{em}\langle\hat{w}_m^{(t)}-c_m,\mathbb{E}(X_e^TX_e)(\theta_{em}^{(t)}-\mu_e)\rangle\\ &= \|\hat{w}_m^{(t)}-c_m\|^2+4\eta_t^2\delta^2\sum_e\pi_{em}\|\theta_{em}^{(t)}-\mu_e\|^2-\underbrace{4\eta_t\langle\hat{w}_m^{(t)}-c_m,\sum_e\pi_{em}\delta^2(\theta_{em}^{(t)}-\mu_e)\rangle}_{C_1}. \qquad (59)\end{aligned}$$

$$C_1 = -4\eta_t \sum_e \pi_{em}\langle \hat{w}_m^{(t)} - \theta_{em}^{(t)}, \delta^2(\theta_{em}^{(t)} - \mu_e)\rangle - 4\eta_t \sum_e \pi_{em}\langle \theta_{em}^{(t)} - c_m, \delta^2(\theta_{em}^{(t)} - \mu_e)\rangle \tag{60}$$

$$\begin{aligned}&\le 4\sum_e \pi_{em}\|\hat{w}_m^{(t)} - \theta_{em}^{(t)}\|^2 + 4\delta^4\eta_t^2 \sum_e \pi_{em}\|\theta_{em}^{(t)} - \mu_e\|^2 - 4\eta_t\delta^2 \sum_e \pi_{em}\|\theta_{em}^{(t)} - \mu_e\|^2 \\ &\qquad - 4\eta_t\delta^2 \underbrace{\sum_e \pi_{em}\langle \mu_e - c_m, \theta_{em}^{(t)} - \mu_e\rangle}_{C_2}\end{aligned} \tag{61}$$

Since $\eta_t \le \frac{1}{4\delta^2}$,

$$\mathbb{E}\|\hat{w}_m^{(t)} - c_m - \eta_t \hat{g}_t\|^2 \tag{62}$$

$$\le \mathbb{E}\|\hat{w}_m^{(t)} - c_m\|^2 + (8\delta^4\eta_t^2 - 4\eta_t\delta^2)\sum_e \pi_{em}\mathbb{E}\|\theta_{em}^{(t)} - \mu_e\|^2 + 4\sum_e \pi_{em}\mathbb{E}\|\hat{w}_m^{(t)} - \theta_{em}^{(t)}\|^2 + C_2 \tag{63}$$

$$\le \mathbb{E}\|\hat{w}_m^{(t)} - c_m\|^2 - 2\eta_t\delta^2 \sum_e \pi_{em}\mathbb{E}\|\theta_{em}^{(t)} - \mu_e\|^2 + 4\sum_e \pi_{em}\mathbb{E}\|\hat{w}_m^{(t)} - \theta_{em}^{(t)}\|^2 + C_2 \tag{64}$$

Note that

$$\sum_e \pi_{em}\mathbb{E}\|\theta_{em}^{(t)} - \mu_e\|^2 \tag{65}$$

$$= \sum_{e\in S_m} \pi_{em}\mathbb{E}\|\theta_{em}^{(t)} - \mu_e\|^2 + \sum_{e\notin S_m} \pi_{em}\mathbb{E}\|\theta_{em}^{(t)} - \mu_e\|^2 \tag{66}$$

$$\ge \sum_{e\in S_m} \pi_{em}(\mathbb{E}\|\theta_{em}^{(t)} - c_m\|^2 + 2r + r^2) + \sum_{e\notin S_m} \pi_{em}\mathbb{E}\|\theta_{em}^{(t)} - \mu_e\|^2 \tag{67}$$

$$= \sum_{e\in S_m} \pi_{em}(\mathbb{E}\|\hat{w}_m^{(t)} - c_m\|^2 + \mathbb{E}\|\hat{w}_m^{(t)} - \theta_{em}^{(t)}\|^2 + 2r + r^2) + \sum_{e\notin S_m} \pi_{em}\mathbb{E}\|\theta_{em}^{(t)} - \mu_e\|^2 \tag{68}$$

And since $\hat{w}_m^{(t)} = \mathbb{E}\sum_e \pi_{em}\theta_{em}^{(t)}$, we have

$$4\mathbb{E}\sum_e \pi_{em}\|\hat{w}_m^{(t)} - \theta_{em}^{(t)}\|^2 \le 4\mathbb{E}\sum_e \pi_{em}\|\hat{w}_m^{(0)} - \theta_{em}^{(t)}\|^2 \tag{69}$$

$$\le 4\sum_e \pi_{em}(T-1)\mathbb{E}\sum_{t'}^{t-1} \eta_t'^2 \|\frac{2}{n_e}X_e^T X_e(\theta_{em}^{(t)} - \mu_e)\|^2 \tag{70}$$

$$\le 16\eta_t^2 E(T-1)^2\delta^4. \tag{71}$$

Thus,

$$\begin{aligned}\mathbb{E}\|\hat{w}_m^{(t)} - c_m - \eta_t\hat{g}_t\|^2 &\le (1 - 2\eta_t\delta^2\sum_e \pi_{em})\mathbb{E}\|\hat{w}_m^{(t)} - c_m\|^2 + 16\eta_t^2 E(T-1)^2\delta^4 \\ &\qquad \underbrace{-2\eta_t\delta^2 \sum_{e\notin S_m} \pi_{em}\mathbb{E}\|\theta_{em}^{(t)} - \mu_e\|^2 - 4\eta_t\delta^2 \sum_e \pi_{em}\langle \theta_{em}^{(t)} - \mu_e, \mu_e - c_m\rangle}_{C_3}\end{aligned} \tag{72}$$

Since

$$C_3 \le 2\eta_t\delta^2 \sum_{\notin S_m} \pi_{em}\|\mu_e - c_m\|^2 - 4\eta_t\delta^2 \sum_{e\in S_m} \pi_{em}\|\theta_{em}^{(t)} + \mu_e\|_2\|\mu_e - c_m\|_2 \tag{73}$$

$$\le 2\eta_t\delta^2 E(1-p) + 4\eta_t\delta^2\gamma_m r \tag{74}$$

we have

$$\mathbb{E}\|\hat{w}_m^{(t)} - c_m - \eta_t \hat{g}_t\|^2 \leq (2\eta_t\delta^2\gamma_m p)\mathbb{E}\|\hat{w}_m^{(t)} - c_m\|^2 + 16\eta_t^2 E(T-1)^2\delta^4 + 2\eta_t\delta^2 E(1-p) + 4\eta_t\delta^2\gamma_m r \tag{75}$$

Notice that

$$\begin{aligned} \mathbb{E}\|\hat{g}_t - g_t\|^2 &= \mathbb{E}\sum_e \frac{4}{n_e^2}\pi_{em}\|(X_e^T X_e - \mathbb{E}(X_e^T X_e))(\theta_{em}^{(t)} - \mu_e)\|^2 + \mathbb{E}\sum_e \frac{4}{n_e^2}\sum_e \pi_{em}\|X_e^T \epsilon_e\|^2 \\ &= E\frac{O(dn_e)}{n_e^2}\delta^4 + E\frac{O(dn_e)}{n_e^2}\delta^2\sigma^2 \end{aligned} \tag{76}$$

so

$$\mathbb{E}\|\hat{w}_m^{(t+1)} - c_m\|_2^2 \leq (1 - 2\eta_t\gamma_m p\delta^2)\mathbb{E}\|\hat{w}_m^{(t)} - c_m\|_2^2 + \eta_t A_1 + \eta_t^2 A_2 \tag{77}$$

where

$$A_1 = 4\delta^2\gamma_m r + 2\delta^2 E(1-p) \tag{78}$$

and

$$A_2 = 16E(T-1)^2\delta^4 + O(\frac{d}{n_e})E(\delta^4 + \delta^2\sigma^2). \tag{79}$$

□

A.2 Convergence of Models with Smooth and Strongly Convex Losses (Theorem 2)

Here we present the detailed proof for Theorem 2.

A.2.1 Key Lemmas

We first state two lemmas for E-step updates and M-step updates, respectively. The proofs of both lemmas are deferred to the Appendix A.2.3

Lemma 3. *Suppose the loss function $\mathcal{L}_{P_t}(\theta)$ is L-smooth and μ-strongly convex for any cluster m. If $\|w_m^{(k)} - w_m^*\| \leq \frac{\sqrt{\mu}R}{\sqrt{\mu}+\sqrt{L}} - r - \Delta$ for some $\Delta > 0$, then E-step updates as*

$$\pi_{em}^{(k)} \geq \frac{\pi_{em}^{(k)}}{\pi_{em}^{(k)} + (1 - \pi_{em}^{(k)})\exp(-\mu R\Delta)}. \tag{80}$$

For M-steps, the local agents are initialized with $\theta_{em}^{(0)} = w_m^{(k)}$. Then for $t = 1, \ldots, T-1$, each agent use local SGD to update its personal model:

$$\theta_{em}^{(t+1)} = \theta_{em} - \eta_t g_{em}(\theta_{em}) = \theta_{em}^{(t)} - \eta_t \nabla \sum_{i=1}^{n_e} \ell(h_{\theta_{em}}(x_e^{(i)}), y_e^{(i)}). \tag{81}$$

To analyze the aggregated model Eq. (6), we define a sequence of virtual aggregated models $\hat{w}_m^{(t)}$.

$$\hat{w}_m^{(t)} = \sum_{e=1}^{E} \frac{\pi_{em}\theta_{em}^{(t)}}{\sum_{e'=1}^{E}\pi_{e'm}}. \tag{82}$$

Lemma 4. *Suppose for any agent $e \in S_m$, its soft clustering label $\pi_{em}^{(k+1)} \geq p$. Then one step local SGD updates $\hat{w}_m^{(t)}$ by Eq.* (83)*, if the learning rate $\eta_t \leq \frac{1}{2(\mu+L)}$.*

$$\mathbb{E}\|\hat{w}_m^{(t+1)} - w_m^*\|_2^2 \leq (1 - \eta_t A_0)\mathbb{E}\|\hat{w}_m^{(t)} - w_m^*\|_2^2 + \eta_t A_1 + \eta_t^2 A_2. \tag{83}$$

where

$$A_0 = \frac{2\gamma_m p\mu L}{\mu + L} \tag{84}$$

$$A_1 = 2\gamma_m L r\sqrt{\frac{2G}{\mu}} + \frac{G(1-p)E}{\mu}(4L + \frac{6}{\mu + L}) + O(r^2). \tag{85}$$

$$A_2 = \frac{4E(T-1)^2 GL^2}{\mu} + \frac{E\sigma^2}{n_e}. \tag{86}$$

Remark. Using this recursive relation, if the learning rate η_t is fixed, the sequence $\hat{w}_m^{(t+1)}$ has a convergence rate of

$$\mathbb{E}\|\hat{w}_m^{(t)} - w_m^*\|^2 \leq (1 - \eta A_0)^t \mathbb{E}\|\hat{w}_m^{(0)} - w_m^*\|^2 + \eta t(A_1 + \eta A_2). \tag{87}$$

Note that the error floor $A_2 = O(T^2)$ where T is the number of local SGD rounds. Therefore, we choose the learning rate $\eta_t \leq \min(\frac{1}{2(\mu+L)}, \frac{\beta}{T^{3/2}})$ so that the error floor is independent from the number of local SGD iterations T.

A.2.2 Completing the Proof of Theorem 2

Theorem 2. *Suppose loss functions have bounded variance for gradients on local datasets, i.e., $\mathbb{E}_{(x,y)\sim\mathcal{D}_e}[\|\nabla\ell(x,y;\theta) - \nabla\mathcal{L}_e(\theta)\|_2^2] \leq \sigma^2$. Assume population losses are bounded, i.e., $\mathcal{L}_e \in G, \forall e \in [E]$. With initialization $\pi_{em}^{(0)} = \frac{1}{M}$ and $\|w_m^{(0)} - w_m^*\| \leq \frac{\sqrt{\mu}R}{\sqrt{\mu}+\sqrt{L}} - r - \Delta_0$ for some $\Delta_0 > 0$, if each agent chooses learning rate $\eta \leq \min(\frac{1}{2(\mu+L)}, \frac{\beta}{T^{3/2}})$, the weights (Π, W) converges by*

$$\pi_{em}^{(k)} \geq \frac{1}{1 + (M-1)\exp(-\mu R \Delta_0 n)}, \ \forall t \in S_m \tag{88}$$

$$\mathbb{E}\|w_m^{(k)} - w_m^*\|^2 \leq (1 - \eta A)^{kT}(\|w_m^{(0)} - w_m^*\|^2 + B) + \frac{\eta AC}{1 - (1 - \eta A)^T}. \tag{89}$$

where k is the total number of communication rounds, and

$$A = \frac{2\gamma_m}{M}\frac{\mu L}{\mu + L}, B = \frac{G(M-1)TE(\frac{4L}{\mu} + \frac{6}{\mu(\mu+L)})}{(1 - \eta A)^T - \exp(-\mu R \Delta_0)}, \tag{90}$$

$$C = \eta^{1/3}\beta^{2/3}(2\gamma_m Lr\sqrt{\frac{2G}{\mu}} + O(r^2)) + \frac{4EGL^2\beta^2}{\mu} + \eta^{4/3}\beta^{2/3}\frac{E\sigma^2}{n_e}. \tag{91}$$

Proof. The proof is quite similar to Theorem 5 for linear models: we follow an induction proof using Lemma 4 and Lemma 5. Suppose Eq. (88) hold for step n. Then for any $t \in S_m$,

$$\pi_{em}^{(k+1)} \geq \frac{\pi_{em}^{(k)}}{\pi_{em}^{(k)} + (1 - \pi_{em}^{(k)})\exp(-\mu R \Delta_n)} \tag{92}$$

$$\geq \frac{1}{1 + (M-1)\exp(-\mu R \Delta_0 n)\exp(-\mu R \Delta_n)} \tag{93}$$

$$\geq \frac{1}{1 + (M-1)\exp(-\mu R \Delta_0 (k+1))} \tag{94}$$

We recall the virtual sequence $\hat{w}_m^{(t)}$ defined in Eq. (82). Models are synchronized after T rounds of local iterations, so $w_m^{(k+1)} = \hat{w}_m^{(T)}$. Thus, according to Lemma 4,

$$\mathbb{E}\|w_m^{(k+1)} - w_m^*\|^2 = \mathbb{E}\|\hat{w}_m^{(T)} - w_m^*\|^2 \tag{95}$$

$$\leq (1 - \eta A_0)^T \mathbb{E}\|w_m^{(k)} - w_m^*\|^2 + \eta T(A_1 + \eta A_2) \tag{96}$$

$$\leq (1 - \eta A_0)^T \Big((1 - \eta A)^{kT}(\mathbb{E}\|w_m^{(0)} - w_m^*\|^2) + B((1 - \eta A)^{kT} - \exp(-\mu R \Delta_0 k)) + \frac{\eta AC}{1 - (1 - \eta A)^T}\Big) + \eta T(A_1 + \eta A_2) \tag{97}$$

$$\leq (1 - \eta A)^{(k+1)T}\mathbb{E}\|w_m^{(0)} - w_m^*\|^2 + \underbrace{(1 - \eta A)^T B\big((1 - \eta A)^{kT} - \exp(-\mu R \Delta_0 k)\big) + \eta\frac{GT(1-p)E}{\mu}(4L + \frac{6}{\mu + L})}_{F_1}$$

$$+ \underbrace{(1 - \eta A)^T \frac{\eta AC}{1 - (1 - \eta A)^T} + \eta T(2\gamma_m Lr\sqrt{\frac{2G}{\mu}} + O(r^2)) + \eta^2 T A_2}_{F_2}. \tag{98}$$

For F_1, we use the fact that

$$\pi_{em}^{(k+1)} \geq \frac{1}{1 + (M-1)\exp-(\mu R \Delta_0 (k+1))} \geq 1 - (M-1)\exp(-\mu R \Delta(N-1)),$$

so

$$F_1 \leq (1-\eta A)^T B\big((1-\eta A)^{kT} - \exp(-\mu R\Delta_0 n)\big) + \eta\frac{G(M-1)\exp(-\mu R\Delta_0 n)}{\mu}(4L + \frac{6}{\mu+L}) \tag{99}$$

$$= B\Big((1-\eta A)^{(k+1)T} - \exp(-\mu R\Delta_0 n)\Big) \tag{100}$$

For F_2, since $\eta \leq \frac{\beta}{T^{3/2}}$, we have

$$F_2 \leq (1-\eta A)^T \frac{\eta AC}{1-(1-\eta A)^T} + \eta^{1/3}\beta^{2/3}(2\gamma_m Lr\sqrt{\frac{2G}{\mu}} + O(r^2)) + \frac{4EGL^2\beta^2}{\mu} + \eta^{4/3}\beta^{2/3}\frac{E\sigma^2}{n_e} \tag{101}$$

$$= \frac{\eta AC}{1-(1-\eta A)^T}. \tag{102}$$

Combining F_1 and F_2 finishes the induction for Eq. (89). □

A.2.3 Deferred Proofs of Key Lemmas

Lemma 3.

Proof. According to Algorithm 1,

$$\pi_{em}^{(k+1)} = \frac{\pi_{em}^{(k)}}{\pi_{em}^{(k)} + \sum_{m'\neq m}\pi_{em'}^{(k)}\exp\Big(\mathbb{E}\ell(x,y;w_m^n) - \mathbb{E}\ell(x,y;w_{m'}^n)\Big)} \tag{103}$$

$$\geq \frac{\pi_{em}^{(k)}}{\pi_{em}^{(k)} + (1-\pi_{em}^{(k)})\exp\Big(\max_{m'\neq m}(\mathcal{L}_{P_t}(w_m^{(k)}) - \mathcal{L}_{P_t}(w_{m'}^{(k)}))\Big)} \tag{104}$$

Since $\mathcal{L}_{P_t}$ is L-smooth and μ-strongly convex,

$$\begin{aligned}\mathcal{L}_{P_t}(w_m^{(k)}) - \mathcal{L}_{P_t}(w_{m'}^{(k)}) &\leq \frac{L}{2}\|w_m^{(k)} - \theta_t^*\|^2 - \frac{\mu}{2}\|w_{m'}^{(k)} - \theta_t^*\|^2 \\ &\leq \frac{L}{2}(\frac{\sqrt{\mu}R}{\sqrt{\mu}+\sqrt{L}} - \Delta)^2 - \frac{\mu}{2}(\frac{\sqrt{L}R}{\sqrt{\mu}+\sqrt{L}} + \Delta)^2 \\ &\leq -\sqrt{\mu L}R\Delta + \frac{L-\mu}{2}\Delta^2 \leq -\mu R\Delta. \end{aligned} \tag{105}$$

Combining Eq. (104) and Eq. (105) completes our proof. □

Lemma 4.

Proof. We define $g_m^{(t)} = \sum_e \pi_{em}\frac{1}{n_e}\sum_{i=1}^{n_e}\nabla\ell(h_{\theta_{em}}(x_e^{(i)}), y_e^{(i)})$ and $\hat{g}_m^{(t)} = \sum_e \pi_{em}\nabla\mathcal{L}(\theta_{em}^{(t)})$.

$$\mathbb{E}\|\hat{w}_m^{(t+1)} - w_m^*\|^2 = \mathbb{E}\|\hat{w}_m^{(t)} - w_m^* - \eta_t g_m\|^2 \tag{106}$$

$$= \mathbb{E}\|\hat{w}_m^{(t)} - w_m^* - \eta_t\hat{g}_m^{(t)}\|^2 + \eta_t^2\mathbb{E}\|g_m^{(t)} - \hat{g}_m^{(t)}\|^2 + 2\eta_t\mathbb{E}\langle w_m^{(t)} - w_m^* - \eta_t\hat{g}_m^{(t)}, \hat{g}_m^{(t)} - g_m^{(t)}\rangle \tag{107}$$

$$= \mathbb{E}\|\hat{w}_m^{(t)} - w_m^* - \eta_t\hat{g}_m^{(t)}\|^2 + \eta_t^2\mathbb{E}\|g_m^{(t)} - \hat{g}_m^{(t)}\|^2. \tag{108}$$

The first term can be decomposed into

$$\|\hat{w}_m^{(t)} - w_m^* - \eta_t\hat{g}_m^{(t)}\|^2 = \|\hat{w}_m^{(t)} - w_m^*\|^2 + \eta_t^2\|\hat{g}_m^{(t)}\|^2 - 2\eta_t\langle\hat{w}_m^{(t)} - w_m^*, \hat{g}_m^{(t)}\rangle. \tag{109}$$

Note that

$$\|\hat{g}_m^{(t)}\|^2 \leq \sum_{e=1}^{E} \pi_{em} \|\nabla \mathcal{L}_e(\theta_{em}^{(t)})\|^2. \tag{110}$$

$$-\langle \hat{w}_m^{(t)} - w_m^*, \hat{g}_m^{(t)} \rangle = -\sum_{e=1}^{E} \pi_{em} \langle \hat{w}_m^{(t)} - \theta_{em}(t), \nabla \mathcal{L}_e(\theta_{em}^{(t)}) \rangle - \sum_{e=1}^{E} \pi_{em} \langle \theta_{em}^{(t)} - w_m^*, \nabla \mathcal{L}_e(\theta_{em}^{(t)}) \rangle. \tag{111}$$

We further decompose the two terms in Eq. (111) by

$$-2\langle \hat{w}_m^{(t)} - \theta_{em}^{(t)}, \nabla \mathcal{L}_e(\theta_{em}^{(t)}) \rangle \leq \frac{1}{\eta_t} \|\hat{w}_m^{(t)} - \theta_{em}^{(t)}\|^2 + \eta_t \|\nabla \mathcal{L}_e(\theta_{em}^{(t)})\|^2. \tag{112}$$

and

$$\langle \theta_{em}^{(t)} - w_m^*, \nabla \mathcal{L}_e(\theta_{em}^{(t)}) \rangle \geq \langle \theta_{em}^{(t)} - w_m^*, \nabla \mathcal{L}_e(\theta_{em}^{(t)}) - \nabla \mathcal{L}_e(w_m^*) \rangle + \|\nabla \mathcal{L}_e(w_m^*)\|_2 \|\theta_{em}^{(t)} - w_m^*\|_2. \tag{113}$$

$$\geq \frac{\mu L}{\mu + L} \|\theta_{em}^{(t)} - w_m^*\|^2 + \frac{1}{\mu + L} \|\nabla \mathcal{L}_e(\theta_{em}^{(t)} - \nabla \mathcal{L}_e(w_m^*))\|^2 + \|\nabla \mathcal{L}_e(w_m^*)\|_2 \|\theta_{em}^{(t)} - w_m^*\|_2. \tag{114}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\|\hat{w}_m^{(t+1)} - w_m^*\|^2 = {} & \underbrace{\mathbb{E}\|\hat{w}_t - w_m^*\|^2 - 2\eta_t \frac{\mu L}{\mu + L} \sum_e \pi_{em} \mathbb{E}\|\theta_{em}^{(t)} - w_m^*\|^2}_{E_1} + \underbrace{\sum_e \pi_{em} \mathbb{E}\|\hat{w}_m^{(t)} - \theta_{em}^{(t)}\|^2}_{E_2} \\
& + \underbrace{\Big(2\eta_t^2 \sum_e \pi_{em} \mathbb{E}\|\nabla \mathcal{L}_e(\theta_{em}^{(t)})\|^2 - 2\eta_t \frac{1}{\mu + L} \sum_e \pi_{em} \mathbb{E}\|\nabla \mathcal{L}_e(\theta_{em}^{(t)}) - \nabla \mathcal{L}_e(w_m^*)\|^2\Big)}_{E_3} \\
& + \underbrace{2\eta_t \mathbb{E} \sum_e \pi_{em} \|\theta_{em}^{(t)} - w_m^*\|_2 \cdot \|\nabla \mathcal{L}_e(w_m^*)\|_2}_{E_4} + \underbrace{\eta_t^2 \mathbb{E}\|g_m^{(t)} - \hat{g}_m^{(t)}\|^2}_{E_5}.
\end{aligned} \tag{115}$$

□

$$\begin{aligned}
E_1 &= \mathbb{E}\|\hat{w}_t - w_m^*\|^2 - 2\eta_t \frac{\mu L}{\mu + L} \mathbb{E}\Big(\sum_e \pi_{em} \|\hat{w}_m^{(t)} - w_m^*\|^2 + \sum_e \pi_{em} \|\hat{w}_m^{(t)} - \theta_{em}^{(t)}\|^2\Big) \\
&\leq (1 - \frac{2\eta_t \mu L p \gamma_m}{\mu + L}) \mathbb{E}\|w_m^{(t)} - w_m^*\|^2 + E_2.
\end{aligned} \tag{116}$$

$$\begin{aligned}
E_2 &= \mathbb{E} \sum_e \pi_{em} \|\hat{w}_m^{(t)} - \theta_{em}^{(t)}\|^2 \\
&= \mathbb{E} \sum_e \pi_{em} \|(w_m^{(0)} - \theta_{em}^{(t)}) + (\theta_{em}^{(t)} - w_m^{(t)})\|^2 \\
&\leq \mathbb{E} \sum_e \pi_{em} \|(w_m^{(0)} - \theta_{em}^{(t)})\|^2 \\
&\leq \sum_e \pi_{em} (T-1) \mathbb{E} \sum_{t'=0}^{t-1} {\eta_{t'}}^2 \|g_{em}(\theta_{em}^{(t')})\|^2 \\
&\leq \frac{2\eta_t^2 E (T-1)^2 G^2 L^2}{\mu}.
\end{aligned} \tag{117}$$

$$
\begin{aligned}
E_3 &= 2\mathbb{E}\sum_e \pi_{em}\Big((\eta_t^2 - \frac{\eta_t}{\mu+L})\|\nabla\mathcal{L}_e(\theta_{em}^{(t)})\|^2 + \frac{2\eta_t}{\mu+L}\langle\nabla\mathcal{L}_e(\theta_{em}^{(t)}), \nabla\mathcal{L}_e(w_m^*)\rangle - \eta_t\frac{\|\nabla\mathcal{L}_e(w_m^*)\|^2}{\mu+L}\Big) \\
&\le 2\eta_t\mathbb{E}\sum_e \pi_{em}\Big(\frac{1}{2(\mu+L)}\|\nabla\mathcal{L}_e(\theta_{em}^{(t)})\|^2 + \frac{1}{\mu+L}\langle\nabla\mathcal{L}_e(\theta_{em}^{(t)}), \nabla\mathcal{L}_e(w_m^*)\rangle - \frac{\|\nabla\mathcal{L}_e(\theta_{em}^{(t)})\|^2}{\mu+L}\Big) \\
&\le 6\eta_t\mathbb{E}\frac{\|\nabla\mathcal{L}_e(w_m^*)\|^2}{\mu+L} \\
&\le 6\eta_t\sum_{e\in S_m}\pi_{em}\frac{L^2r^2}{\mu+L} + 6\eta_t\sum_{e\notin S_m}\pi_{em}\frac{2G}{\mu(\mu+L)} \\
&\le \eta_t O(r^2) + 6\eta_t\frac{G(1-p)E}{\mu(\mu+L)}. \qquad (118)
\end{aligned}
$$

$$
\begin{aligned}
E_4 &= 2\eta_t\mathbb{E}\sum_{e\in S_m}\pi_{em}\|\theta_{em}^{(t)} - w_m^*\|_2\cdot\|\nabla\mathcal{L}_e(w_m^*)\|_2 + 2\eta_t\mathbb{E}\sum_{e\notin S_m}\pi_{em}\|\theta_{em}^{(t)} - w_m^*\|_2\cdot\|\nabla\mathcal{L}_e(w_m^*)\|_2 \\
&\le 2\eta_t\gamma_m Lr\sqrt{\frac{2G}{\mu}} + 2\eta_t(1-p)EL\cdot\frac{2G}{\mu}. \qquad (119)
\end{aligned}
$$

$$
\begin{aligned}
E_5 &= \eta_t^2\mathbb{E}\|g_m^{(t)} - \hat{g}_m^{(t)}\|^2 \\
&\le \eta_t^2\mathbb{E}\Big\|\sum_e\pi_{em}\Big(\frac{1}{n_e}\sum_{i=1}^{n_e}\nabla\ell(h_{\theta_{em}}(x_e^{(i)}), y_e^{(i)}) - \mathcal{L}(\theta_{em}^{(t)})\Big)\Big\|^2 \\
&\le \eta_t^2 E\frac{\sigma^2}{n_e}. \qquad (120)
\end{aligned}
$$

Combining Eq. (116) to Eq. (120) yields the conclusion of Lemma 4.

B Fairness Analysis

B.1 Proof of Theorem 3

Proof. Let the first cluster m_1 contain agents $\mu_1, \dots, \mu_{E-1}$, while the second cluster contains only the outlier μ_E. Then, for $e = 1, \dots, E-1$,

$$
\mathcal{E}_e(w_{m_1}) = \delta^2\left\|\mu_e - \frac{\sum_{e'=1}^{E-1}\mu_{e'}}{E-1}\right\|^2 \le \delta^2 r^2 \qquad (121)
$$

And for the outlier agent, the expected output is just the optimal solution, so

$$
\mathcal{E}_E(w_{m_2}) = 0 \qquad (122)
$$

As a result, the fairness of this algorithm is bounded by

$$
\mathcal{F}_M(P) = \max_{i,j\in[E]}|\mathcal{E}_i(\Pi, W) - \mathcal{E}_j(\Pi, W)| \le \delta^2 r^2. \qquad (123)
$$

On the other hand, the expected final weights of of FedAvg algorithm is $w_{avg} = \bar{\mu} = \frac{\sum_{e=1}^E \mu_e}{E}$, so the expected loss for agent e shall be

$$
\mathbb{E}_{(x,y)\sim\mathcal{P}_e}(\ell_{\hat{\theta}}(x)) = \mathbb{E}_{x\sim\mathcal{N}(0,\delta^2 I_d),\epsilon\sim\mathcal{N}(0,\sigma^2)}[(\mu_i^T x + \epsilon - \bar{\mu}^T x)^2] = \sigma^2 + \delta^2\|\mu_e - \bar{\mu}\|^2 \qquad (124)
$$

The infimum risk for agent t_1 is σ_1^2, and after subtracting it from the expected loss, we have

$$\begin{aligned}
\mathcal{E}_1(w_{avg}) &= \delta^2\|\mu_1 - \bar{\mu}\|^2 \\
&= \delta^2\|\mu_1 - \frac{\sum_{e=1}^{E-1}\mu_1}{E} - \frac{\mu_E}{E}\|^2 \\
&\leq \delta^2\Big(r \cdot \frac{E-1}{E} + \frac{\|\mu_1 - \mu_E\|}{E}\Big)^2 \\
&\leq \delta^2(r \cdot \frac{E-1}{E} + \frac{R+r}{E})^2 = \delta^2(r + \frac{R}{E})^2 \qquad (125)
\end{aligned}$$

However for the outlier agent,

$$\mathcal{E}_{P_E}(w_{avg}) = \delta^2\|\mu_E - \bar{\mu}\|^2 \qquad (126)$$

$$= \delta^2\left\|\frac{E-1}{E}\mu_E - \frac{\sum_{e=1}^{E-1}\mu_E}{E}\right\|^2 \qquad (127)$$

$$\geq \Big(\frac{E-1}{E}\Big)^2\delta^2 R^2 \qquad (128)$$

Hence,

$$\mathcal{F}_{avg}(P) \geq \mathcal{E}_E(w_{avg}) - \mathcal{E}_1(w_{avg}) = \delta^2\Big(\frac{R^2(E-2) - 2Rr}{E} + r^2\Big) \qquad (129)$$

□

B.2 Proof of Theorem 4

Proof. Note that the local population loss for agent i with weights θ is

$$\mathcal{L}_i(\theta) = \int p_i(x,y)\ell(f_\theta(x), y)\mathrm{d}x\mathrm{d}y. \qquad (130)$$

Thus,

$$|\mathcal{L}_i(\theta_i^*) - \mathcal{L}_j(\theta_i^*)| = \int |p_i(x,y) - p_j(x,y)| \cdot \ell(f_{\theta_i^*}(x), y)\mathrm{d}x\mathrm{d}y \qquad (131)$$

$$\leq G \cdot \int |p_i(x,y) - p_j(x,y)|\mathrm{d}x\mathrm{d}y \leq Gr. \qquad (132)$$

Hence,

$$\mathcal{L}_i(\theta_j^*) \leq \mathcal{L}_j(\theta_j^*) + Gr \leq \mathcal{L}_j(\theta_i^*) + Gr \leq \mathcal{L}_i(\theta_i^*) + 2Gr. \qquad (133)$$

For the cluster that combines agents $\{1, \ldots, E-1\}$ together, the weight converges to $\bar{\theta}' = \frac{1}{E-1}\sum_{i=1}^{E-1}\theta_i^*$. Then $\forall i = 1, \ldots, E-1$, the population loss for the ensemble prediction

$$\mathcal{L}_i(\theta, \Pi) = \mathcal{L}_i\Big(\frac{\sum_{j=1}^{E-1}\theta_j^*}{E-1}\Big) \qquad (134)$$

$$\leq \frac{1}{T-1}\sum_{j=1}^{T-1}\mathcal{L}_i(\theta_j^*) \qquad (135)$$

$$\leq \mathcal{L}_i(\theta_i^*) + \frac{2Gr}{E-1}. \qquad (136)$$

Therefore, for any $i = 1, \ldots, T-1$,

$$\mathcal{E}_i(\theta, \Pi) \leq \frac{2Gr}{E-1}. \qquad (137)$$

Since $\mathcal{E}_T(\theta, \Pi) = 0$,

$$\mathcal{F}_{EM}(W, \Pi) \leq \frac{2Gr}{E-1} \qquad (138)$$

Now we prove the second part of Theorem 4 for the fairness of Fedavg algorithm. For simplicity, we define $B = \frac{2Gr}{E-1}$ in this proof. Also, we denote the mean of all optimal weight $\bar{\theta} = \frac{\sum_{i=1}^{E} \theta_i^*}{E}$ and $\bar{\theta}' = \frac{\sum_{i=1}^{E-1} \theta_i^*}{E-1}$.

Remember that we assume loss functions to be L-smooth, so

$$\mathcal{L}_E(\theta_i^*) \leq \mathcal{L}_E(\bar{\theta}') + \langle \nabla \mathcal{L}_E(\bar{\theta}'), \theta_i^* - \bar{\theta}' \rangle + \frac{L}{2}\|\bar{\theta}' - \theta_i\|^2. \tag{139}$$

Taking summation over $i = 1, \ldots, E-1$, we get

$$\begin{aligned}
\mathcal{L}_E(\bar{\theta}') &\geq \frac{1}{E-1}\Big(\sum_{i=1}^{E-1} \mathcal{L}_E(\theta_i^*) - \langle \nabla \mathcal{L}_E(\bar{\theta}'), \sum_{i=1}^{E-1}(\theta_i - \bar{\theta}')\rangle - \frac{L}{2}\sum_{i=1}^{E-1}\|\bar{\theta}' - \theta_i\|^2\Big) & (140)\\
&= \frac{1}{E-1}\Big(\sum_{i=1}^{E-1} \mathcal{L}_E(\theta_i^*) - \frac{L}{2}\sum_{i=1}^{E-1}\|\bar{\theta}' - \theta_i\|^2\Big) & (141)\\
&\geq \mathcal{L}_E(\theta_E^*) + R - \frac{LB}{\mu}. & (142)
\end{aligned}$$

The last inequality uses the μ-strongly convex condition that implies

$$B \geq \mathcal{L}_i(\bar{\theta}') - \mathcal{L}_i(\theta_i^*) \geq \frac{\mu}{2}\|\bar{\theta}' - \theta_i\|^2. \tag{143}$$

By L-smoothness, we have

$$\mathcal{L}_E(\bar{\theta}') \leq \mathcal{L}_E(\bar{\theta}) + \langle \nabla \mathcal{L}_E(\bar{\theta}), \bar{\theta}' - \bar{\theta} \rangle + \frac{L}{2}\|\bar{\theta}' - \bar{\theta}\|^2. \tag{144}$$

$$\mathcal{L}_E(\theta_E^*) \leq \mathcal{L}_E(\bar{\theta}) + \langle \nabla \mathcal{L}_E(\bar{\theta}), \theta_E^* - \bar{\theta} \rangle + \frac{L}{2}\|\theta_E^* - \bar{\theta}\|^2. \tag{145}$$

Note that $\bar{\theta} = \frac{\bar{\theta}' + (E-1)\theta_E^*}{E}$, we take a weighted sum over the above two inequalities to cancel the dot product terms out. We thus derive

$$\begin{aligned}
\mathcal{L}_E(\bar{\theta}) &\geq \frac{(E-1)\mathcal{L}_E(\bar{\theta}') + \mathcal{L}_E(\theta_E^*) - \frac{L}{2}(E-1)\|\bar{\theta}' - \bar{\theta}\|^2 - \frac{L}{2}\|\theta_E^* - \bar{\theta}\|^2}{E} & (146)\\
&= \frac{E-1}{E}\Big(R - \frac{LB}{\mu} - \frac{L\|\theta_E^* - \bar{\theta}'\|^2}{2E}\Big) + \mathcal{L}_E(\theta_E^*). & (147)
\end{aligned}$$

Note that $\mathcal{L}_E(\cdot)$ is μ-strongly convex, which means

$$R - \frac{LB}{\mu} \geq \mathcal{L}_E(\bar{\theta}') - \mathcal{L}_E(\theta_E^*) \geq \frac{\mu}{2}\|\theta_E^* - \bar{\theta}'\|^2. \tag{148}$$

so

$$\mathcal{L}_E(\bar{\theta}) \geq (1 - \frac{L}{\mu E}) \cdot \frac{E-1}{E}(R - \frac{LB}{\mu}) + \mathcal{L}_E(\theta_E^*). \tag{149}$$

And

$$\mathcal{E}_E(\bar{\theta}) \geq (1 - \frac{L}{\mu E}) \cdot \frac{E-1}{E}(R - \frac{LB}{\mu}). \tag{150}$$

On the other hand, for agent $i = 1, \ldots, E-1$ we know

$$\mathcal{L}_i(\bar{\theta}) \leq \mathcal{L}_i(\bar{\theta}') + \langle \nabla \mathcal{L}_i(\bar{\theta}'), \bar{\theta} - \bar{\theta}' \rangle + \frac{L}{2}\|\bar{\theta} - \bar{\theta}'\|^2. \tag{151}$$

By L smoothness,

$$\|\nabla \mathcal{L}_i(\bar{\theta}')\|_2 \leq L\|\bar{\theta}' - \theta_i^*\| \leq L\sqrt{\frac{2B}{\mu}}. \tag{152}$$

So

$$\mathcal{L}_i(\bar{\theta}) \leq \mathcal{L}_i(\theta_i^*) + B + L\sqrt{\frac{2B}{\mu}}\sqrt{\frac{2(R-\frac{LB}{\mu})}{\mu}}\frac{1}{E} + \frac{L(R-\frac{LB}{\mu})}{\mu E^2} \tag{153}$$

$$\mathcal{E}_i(\bar{\theta}) \leq B + \frac{2L}{\mu E}\sqrt{B(R-\frac{LB}{\mu})} + \frac{L(R-\frac{LB}{\mu})}{\mu E^2} \tag{154}$$

In conclusion, the fairness can be estimated by

$$\begin{aligned} \mathcal{F}_{avg}(P) &\geq \mathcal{E}_E(\bar{\theta}) - \mathcal{E}_1(\bar{\theta}) && (155) \\ &\geq \Big(\frac{E-1}{E} - \frac{L}{\mu E^2}\Big)R - \Big(1 + \frac{L(E-1)}{\mu E} - \frac{L^2}{\mu^2 E}\Big)B - \frac{2L}{\mu E}\sqrt{B(R-\frac{L}{\mu}B)} && (156) \end{aligned}$$

□

B.3 Proof of Divergence Reduction

Here we prove the claim that the assumption $\mathcal{L}_E(\theta_e^*) - \mathcal{L}_E(\theta_E^*) \geq R$ is implied by a lower bound of the H-divergence [31].

$$D_H(\mathcal{P}_e, \mathcal{P}_E) \geq \frac{LR}{4\mu} \tag{157}$$

Proof. Note that

$$\begin{aligned} D_H(\mathcal{P}_e, \mathcal{P}_E) &= \frac{1}{2}\min_{\theta}\Big(\mathcal{L}_e(\theta) + \mathcal{L}_E(\theta)\Big) + \frac{1}{2}\Big(\mathcal{L}_e(\theta_e^*) + \mathcal{L}_E(\theta_E^*)\Big) && (158) \\ &\leq \frac{1}{2}\Big(\mathcal{L}_e(\frac{\theta_e^* + \theta_E^*}{2}) + \mathcal{L}_E(\frac{\theta_e^* + \theta_E^*}{2})\Big) - \frac{1}{2}\Big(\mathcal{L}_e(\theta_e^*) + \mathcal{L}_E(\theta_E^*)\Big) && (159) \\ &\leq \frac{1}{2} \times (\frac{1}{2}L\|\frac{\theta_E^* - \theta_e^*}{2}\|_2^2 \times 2) && (160) \\ &= \frac{1}{8}L\|\theta_E^* - \theta_e^*\|_2^2 && (161) \end{aligned}$$

Therefore,

$$\begin{aligned} \mathcal{L}_E(\theta_e^*) - \mathcal{L}_E(\theta_E^*) &\geq \frac{\mu\|\theta_E^* - \theta_e^*\|_2^2}{2} && (162) \\ &\geq \frac{\mu}{2}\frac{8D_H(\mathcal{P}_e, \mathcal{P}_E)}{L} = R. && (163) \end{aligned}$$

□

C Experimental Details

Here we elaborate more details of our experiments.

Machines. We simulate the federated learning setup on a Linux machine with AMD Ryzen Threadripper 3990X 64-Core CPUs and 4 NVIDIA GeForce RTX 3090 GPUs.

Hyperparameters. For each FL experiment, we implement both FOCUS algorithm and FedAvg algorithm using SGD optimizer with the same hyperparameter setting. Detailed hyperparameter specifications are listed in Table 2 for different datasets, including learning rate, the number local training steps, batch size, the number of training epochs, etc.

Table 2: Dataset description and hyperparameters.

Dataset	# training samples	# test samples	E	M	batch size	learning rate	local training epochs	epochs
MNIST	60000	10000	10	3	6000	0.1	10	300
CIFAR	50000	10000	10	2	100	0.1	10	200
Yelp/IMDB	56000/25000	38000/25000	10	2	512	5e-5	2	3

D Broader Impact

This paper presents a novel definition of fairness via agent-level awareness for federated learning, which takes the heterogeneity of local data distributions among agents into account. We develop FAA as a fairness metric for Federated learning and design FOCUS algorithm to improve the corresponding fairness. We believe that FAA can benefit the ML community as a standard measurement of fairness for FL based on our theoretical analyses and empirical results.

A possible negative societal impact may come from the misunderstanding of our work. For example, low FAA does not necessarily mean low loss or high accuracy. Additional utility evaluation metrics are required to evaluate the overall performance of different federated learning algorithms. We have tried our best to define our goal and metrics clearly in Section 3; and state all assumptions for our theorems accurately in Section 4 to avoid potential misuse of our framework.