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

Wenda Chu[1]* **Chulin Xie**[2]* **Boxin Wang**[2] **Linyi Li**[2]
Lang Yin[2] **Arash Nourian**[3] **Han Zhao**[2] **Bo Li**[2]
[1] California Institute of Technology [2] University of Illinois Urbana-Champaign [3] Amazon

Abstract

Federated learning (FL) allows agents to jointly train a global model without sharing their local data to protect the privacy of local agents. However, due to the heterogeneous nature of local data, existing definitions of fairness in the context of FL are prone to noisy agents in the network. For instance, existing work usually considers accuracy parity as the fairness metric for different agents, which is not robust under the heterogeneous setting, since it will enforce agents with high-quality data to achieve similar accuracy to those who contribute low-quality data and may discourage the agents with high-quality data from participating in FL. In this work, we propose a formal FL fairness definition, *fairness via agent-awareness* (FAA), which takes the heterogeneity of different agents into account by measuring the data quality with approximated Bayes optimal error. Under FAA, the performance of agents with high-quality data will not be sacrificed just due to the existence of large numbers of agents with low-quality data. In addition, we propose a fair FL training algorithm leveraging agent clustering (FOCUS) to achieve fairness in FL, as measured by FAA and other fairness metrics. Theoretically, we prove the convergence and optimality of FOCUS under mild conditions for both linear and general convex loss functions with bounded smoothness. We also prove that FOCUS always achieves higher fairness in terms of FAA compared with standard FedAvg under both linear and general convex loss functions. Empirically, we show that on four FL datasets, including synthetic data, images, and texts, FOCUS achieves significantly higher fairness in terms of FAA and other fairness metrics, while maintaining competitive prediction accuracy compared with FedAvg and four state-of-the-art fair FL algorithms.

1 Introduction

Federated learning (FL) is emerging as a promising approach to enable scalable intelligence over distributed settings such as mobile networks [14, 24]. Given the wide adoption of FL in medical analysis [1, 33], recommendation systems [3, 28], and personal Internet of Things (IoT) devices [2], it has become a central question on how to ensure the fairness of the trained global model in FL networks before its large-scale deployment by local agents, especially when the data quality/contributions of different agents are different in the heterogeneous setting.

In standard (centralized) ML, fairness is usually defined as a notion of parity of the underlying distributions from different groups given by a protected attribute (e.g., gender, race). Typical definitions include demographic parity [12, 41], equalized odds [15], and accuracy parity [5, 44]. However, it is yet unclear what is the desired notion of fairness in FL. Previous works that explore fairness in FL mainly focus on the demographic disparity of the final trained model regarding the

[*] Equal Contribution. Correspondence to: wchu@caltech.edu, chulinx2@illinois.edu and lbo@illinois.edu.

International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023 (FL@FM-NeurIPS'23). This workshop does not have official proceedings and this paper is non-archival.

protected attributes as in the centralized setting [7, 17] or the accuracy disparity across agents without considering different contributions of agents [10, 22, 29]. Some works have taken into consideration the local data properties [18, 42] and sizes [11] to measure fairness of FL. However, explicit fairness analysis in FL under heterogeneous agent contributions is still lacking. Thus, in this paper, we aim to ask: *What is a desirable notion of fairness in FL with heterogeneous agents? Can we design efficient training algorithms to guarantee the fairness?* Such fairness notion in FL is critical since on one hand, it will be able to take the heterogeneity of different local agents into account so that it is robust to the potential noisy (or low-contribution) nodes in the network; on the other hand, it will also encourage the participation of benign agents who contribute high-quality data to the FL training.

In light of the above questions, in this work, we aim to define and enhance fairness in FL by explicitly considering the heterogeneity of local agents. In particular, for FL trained with standard FedAvg protocol [27], the global model aims to minimize the loss with respect to the global distribution. In practice, some local agents may contribute low-quality data (e.g., free riders), so intuitively it is "unfair" to train the final model regarding such global distribution over all agents, which will sacrifice the performance of agents with high-quality data. In this paper, we define **fairness via agent-awareness in FL (FAA)** as $\mathcal{FAA}(\{\theta_e\}_{e\in[E]}) = \max_{e_1,e_2\in E} \left|\mathcal{E}_{e_1}(\theta_{e_1}) - \mathcal{E}_{e_2}(\theta_{e_2})\right|$, where $\mathcal{E}_e$ is the *excess risk* of an agent $e \in E$ with model parameter θ_e. Technically, the excess risk of each agent is calculated as $\mathcal{E}_e(\theta_e) = \mathcal{L}_e(\theta_e) - \min_{\theta^*} \mathcal{L}_e(\theta^*)$, which stands for the loss of user e evaluated on the FL model θ_e subtracted by the Bayes optimal error of the local data distribution [30]. For each agent, the excess risk measures how close the test loss $\mathcal{E}_e(\theta_e)$ is to the possible smallest error (i.e., Bayes optimal error) on such a data distribution. In other words, lower excess risk $\mathcal{E}_e(\theta_e)$ indicates a *higher gain* for agent by joining the FL training. Notably, reducing FAA enforces the *equity of excess risks* among agents, following the classic philosophy that agents *"do not suffer from scarcity, but inequality of gains"*. Lower FAA indicates stronger fairness. Based on our fairness definition FAA, we then propose an efficient **fair FL algorithm based on agent clustering** (FOCUS) to improve the fairness of FL by minimizing FAA of agents. Specifically, we first cluster the local agents based on their data distributions and then train a model for each cluster. During inference, the final prediction will be the weighted aggregation over the prediction result of each model trained with the corresponding clustered local data. Theoretically, we provide the convergence analysis and fairness analysis for FOCUS; empirically, we extensively evaluate FOCUS on diverse datasets.

Contributions. In summary, we make contributions on both theoretical and empirical fronts.

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

2 Related work

Fair Federated Learning There have been several studies exploring fairness in FL. Li et al. [22] first define agent-level fairness by considering *accuracy equity* across agents and achieve fairness by assigning the agents with worse performance with higher aggregation weight during training. However, such a definition of fairness fails to capture the heterogeneous nature of local agents. Mohri et al. [29] pursue accuracy parity by improving the performance of the worst-performing agent. Wang et al. [36] propose to mitigate conflict gradients from local agents to enhance fairness. Instead of pursuing fairness with one single global model, Li et al. [23] propose to train a personalized model for each agent to achieve accuracy equity for the personalized models. Zhang et al. [42] predefine the agent contribution levels based on an oracle assumption (e.g., data volume, data collection cost, etc.) for fairness optimization, which lacks quantitative measurement in practice. Xu et al. [38] approximate the Shapely Value based on gradient cosine similarity to evaluate agent contribution. However, Zhang et al. [42] point out that Shapely Value may discourage agents with rare data. Here we provide an algorithm to quantitatively measure the contribution of local data based on each agent's excess risk, which will not be affected even if the agent is the minority.

Clustered Federated Learning Clustered FL algorithms are initially designed for multitasking and personalized federated learning, which assumes that agents can be naturally partitioned into clusters [13, 26, 32, 37]. Existing clustering algorithms usually aim to assign each agent to a cluster that provides the lowest loss [13], optimize the clustering center to be close to the local model [37], or cluster agents with similar gradient updates (with respect to, e.g., cosine similarity [32]) to the same cluster. In addition to these hard clustering approaches (i.e., each agent only belongs to one cluster), soft clustering has also been studied [20, 26, 31, 34], which enables the agents to benefit from multiple clusters. However, none of these works considers the fairness of clustered FL and the potential implications, and our work makes the first attempt to bridge them.

3 Fair Federated Learning on Heterogeneous Data

We first define our fairness via agent-awareness in FL with heterogeneous data and then introduce our fair FL based on the agent clustering (FOCUS) algorithm to achieve FAA.

3.1 Fairness via Agent-Awareness in FL (FAA)

Given a set of E agents participating in the FL network, each agent e only has access to its local dataset: $D_e = \{(x_e, y_e)\}_{i=1}^{n_e}$, which is sampled from a distribution $\mathcal{P}_e$. The goal of standard FedAvg training is to minimize the overall loss $\mathcal{L}_E(\theta)$ based on the local loss $\mathcal{L}_e(\theta)$ of each agent: $\min_\theta \mathcal{L}_E(\theta) = \sum_{e\in[E]} \frac{|D_e|}{n} \mathcal{L}_e(\theta)$, where $\mathcal{L}_e(\theta) = \mathbb{E}_{(x,y)\in\mathcal{P}_e} \ell(h_\theta(x), y)$. where $\ell(\cdot,\cdot)$ is a loss function given model prediction $h_\theta(x)$ and label y (e.g., cross-entropy loss), $n = \sum_{e\in[E]} |D_e|$ represents the total number of training samples, and θ represents trained global model.

Intuitively, the performance of agents with high-quality and clean data could be severely compromised by the existence of large amounts of agents with low-quality and noisy data under FedAvg. To solve such a problem and characterize the distinctions of local data distributions (contributions) among agents to ensure fairness, we propose fairness via agent-awareness in FL (FAA) as follows.

Definition 3.1 (Fairness via agent-awareness for FL (FAA)). Given a set of agents E in FL, the overall fairness score among all agents is defined as the maximal difference of excess risks for any pair of agents:

$$\mathcal{FAA}(\{\theta_e\}_{e\in[E]}) = \max_{e_1,e_2\in[E]} \Big|\mathcal{E}_{e_1}(\theta_{e_1}) - \mathcal{E}_{e_2}(\theta_{e_2})\Big|. \tag{1}$$

where θ_e is the local model for agent $e \in [E]$. The excess risk $\mathcal{E}_e(\theta_e)$ for agent e given model θ_e is defined as the difference between the population loss $\mathcal{L}_e(\theta_e)$ and the Bayes optimal error of the corresponding data distribution, i.e., $\mathcal{E}_e(\theta_e) = \mathcal{L}_e(\theta_e) - \min_{\theta^*} \mathcal{L}_e(\theta^*)$, where θ^* denotes any possible models.

Note that in FedAvg, each client uses the global model θ as its local model θ_e. Definition 3.1 represents a quantitative data-dependent measurement of agent-level fairness. Instead of forcing accuracy parity among all agents regardless of their data quality, we define agent-level fairness as the equity of *excess risks* among agents, which takes the contributions of local data into account by measuring their Bayes errors. For instance, when a local agent has low-quality data, although the corresponding utility loss would be high, the Bayes error of such low-quality data is also high, and thus the excess risk of the user is still low, enabling the agents with high-quality data to achieve low utility loss for fairness. When local data distributions are homogeneous, FAA reduces to the fairness definition of agnostic loss [29]. Therefore, FAA is a generalization of agnostic loss that accommodates both homogeneous and heterogeneous data distributions. We defer more detailed discussion to Appendix C.

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

Algorithm 1 EM clustered federated learning algorithm

```
Input: Agents with data $\{D_i\}_{i\in[E]}$ and $M$ models.
Initialize $w_m^{(0)}$ and $\pi_{em}^{(0)} = \frac{1}{M}$ for $m \in [M], e \in [E]$.
for $t = 0$ to $T-1$ do
    for agent $e \in [E]$ do
        for model $m \in [M]$ (E step) do
            $$\pi_{em}^{(t+1)} \leftarrow \frac{\pi_{em}^{(t)} \exp\Big(-\mathbb{E}_{(x,y)\in D_e}\ell(x,y;w_m^{(t)})\Big)}{\sum_{m=1}^{M}\pi_{em}^{(t)} \exp\Big(-\mathbb{E}_{(x,y)\in D_e}\ell(x,y;w_m^{(t)})\Big)} \qquad (2)$$
        end for
    end for
    for model $m \in [M]$ (M step) do
        $$w_m^{(t+1)} \leftarrow \arg\min_w \sum_{e=1}^{E} \pi_{em}^{(t+1)} \sum_{i=1}^{n_e} \ell\Big(h_w(x_e^{(i)}), y_e^{(i)}\Big) \qquad (3)$$
    end for
end for
Return model weights $w_m^{(T)}$
```

Method Overview. To enhance the fairness of FL in terms of FAA, we provide an agent clustering-based FL algorithm, FOCUS (Algorithm 1), by partitioning agents conditioned on their data distributions. Intuitively, grouping agents with similar local data distributions and similar contributions together helps to improve fairness, since it reduces the intra-cluster data heterogeneity. FOCUS leverages the Expectation-Maximization algorithm to perform agent clustering. 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; $W = \{w_m\}_{m\in[M]}$ represent the model weights for M data clusters. Given E agents with datasets $D_1, \ldots, D_E$, FOCUS alternately optimizes Π and W in two steps.

E step. Expectation steps update the cluster labels Π given the current estimation of (Π, W). At k-th communication round, the server broadcasts the M cluster models to all agents. The agents calculate the expected training loss $\mathbb{E}_{(x,y)\in D_e}\ell(x,y;w_m^{(t)})$ for each cluster model $w_m^{(t)}$, $m \in [M]$, and then update the soft clustering labels Π according to Equation (2).

M step. The goal of M steps in Equation (3) 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 Equation (4) $\sim$ Equation (6) to estimate the objective in Equation (3). At t-th communication round, for each cluster model $w_m^{(t)}$ received from server, each agent e first initializes its local model $\theta_{em(0)}^{(t)}$ as $w_m^{(t)}$, and then updates the model using its own dataset. To reduce communication costs, each agent is allowed to run SGD locally for K local steps as shown in Equation (5). After K local steps, each agent sends the updated models $\theta_{em(K)}^{(t)}$ back to the central server, and the server aggregates the models of all agents by a weighted average based on the soft clustering labels $\{\pi_{em}\}$. We provide theoretical analysis for the convergence and optimality of FOCUS under these multiple local updates in Section 4.

$$\text{Clients:} \quad \theta_{em(0)}^{(t)} = w_m^{(t)}. \qquad (4)$$

$$\theta_{em(k+1)}^{(t)} = \theta_{em(k)}^{(t)} - \eta_k \nabla \sum_{i=1}^{n_e} \ell\Big(h_{\theta_{em(k)}^{(t)}}(x_e^{(i)}), y_e^{(i)}\Big), \forall k = 1, \ldots, K-1. \qquad (5)$$

$$\text{Server:} \quad w_m^{(t+1)} = \sum_{e=1}^{E} \frac{\pi_{em}^{(t+1)}\theta_{em(K)}^{(t)}}{\sum_{e'=1}^{E}\pi_{e'm}^{(t+1)}}. \qquad (6)$$

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

4 Theoretical Analysis of FOCUS

In this section, we first present the convergence and optimality guarantees of our FOCUS algorithm; and then prove that it improves the fairness of FL regarding FAA. Our analysis considers linear models 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 to deliver the main idea of our analysis. 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 $\mu_e \in \mathbb{R}^d$, and $(x_e^{(i)}, y_e^{(i)})$ is 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 perturbed 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 D_e is $\mathcal{L}_{emp}(D_e; w) = \frac{1}{n_e} \sum_{i=1}^{n_e} (w^T x_e^{(i)} - y_e^{(i)})^2$. We make the following assumption about the heterogeneous agents.

Assumption 4.1. Suppose there are M predefined vectors $\{w_i^*\}_{i=1}^{M}$, where for any $m_1, m_2 \in [M]$, $m_1 \neq m_2$, $\|w_{m_1}^* - w_{m_2}^*\|_2 \geq R$. A set of agents E satisfy separable distributions if they can be partitioned into M subsets $S_1, \ldots, S_M$ such that, for any agent $e \in S_m$, $\|\mu_e - w_m^*\|_2 \leq r < \frac{R}{2}$.

Assumption 4.1 guarantees that the heterogeneous local data distributions are separable so that an optimal clustering solution exists, in which $\{w_1^*, \ldots, w_M^*\}$ are the centers of clusters. We next present Theorem 4.2 to demonstrate the linear convergence rate to the optimal cluster centers for FOCUS. Detailed proofs can be found in Appendix A.1.

Theorem 4.2. *Assume the agent set E satisfies the separable distributions condition in Assumption 4.1. Given trained M models with $\pi_{em}^{(0)} = \frac{1}{M}$, $\forall e, m$. Under the natural initialization w_m for each model $m \in [M]$, which satisfies $\exists \Delta_0 > 0, \|w_m^{(0)} - w_m^*\|_2 \leq \min_{m' \neq m} \|w_m^{(0)} - w_{m'}^*\|_2 - 2(r + \Delta_0)$ and $|D_e| = O(d)$. If learning rate $\eta \leq \min(\frac{1}{4\delta^2}, \frac{\beta}{\sqrt{T}})$, FOCUS converges by*

$$\pi_{em}^{(T)} \geq \frac{1}{1 + (M-1) \cdot \exp(-2R\delta^2 \Delta_0 T)}, \forall e \in S_m, \tag{7}$$

$$\mathbb{E}\|w_m^{(T)} - w_m^*\|_2^2 \leq (1 - \frac{2\eta\gamma_m\delta^2}{M})^{KT} (\|w_m^{(0)} - w_m^*\|_2^2 + A) + 2MKr + M\delta^2 E\beta T^{-1/2} O(K^3, \sigma^2). \tag{8}$$

where T is the total number of communication rounds; K is the number of local updates in each communication round; $\gamma_m = |S_m|$ is the number of agents in the m-th cluster, and

$$A = \frac{2EK(M-1)\delta^2}{(1 - \frac{2\eta\delta^2\gamma_m}{M})^K - \exp(-2R\delta^2\Delta_0)}. \qquad \text{(caused by initial inaccurate clustering)}$$

Remarks. Theorem 4.2 shows the convergence of parameters (Π, W) to a near-optimal solution. Equation (7) implies that the agents will be *correctly clustered* since π_{em} will converge to 1 as the number of communication rounds K increases. In Equation (8), the first term diminishes exponentially, while the second term $2MKr$ reflects the intra-cluster distribution divergence r. The last term originates from the data heterogeneity among clients across different clusters. Its influence is amplified by the number of local updates ($O(K^3)$) and will also diminish to zero as the number of communication rounds T goes to infinity. Our convergence analysis is conditioned on the natural clustering initialization for model weights $w_m^{(0)}$ towards a corresponding cluster center w_m^*, which is standard in convergence analysis for a mixture of models [4, 39]. Detailed proofs can be found in Appendix A.1.

Smooth and strongly convex loss functions. Next, we extend our analysis to a more general case of non-linear models with L-smooth and μ-strongly convex loss function.

Assumption 4.3 (Smooth and strongly convex loss functions)**.** The population loss functions $\mathcal{L}_e(\theta)$ for each agent e is L-smooth, i.e., $\|\nabla^2 \mathcal{L}_e(\theta)\|_2 \leq L$. The loss functions are μ-strongly convex, if the eigenvalues λ of the Hessian matrix $\nabla^2 \mathcal{L}_e(\theta)$ satisfy $\lambda_{\min}(\nabla^2 \mathcal{L}_e(\theta)) \geq \mu$.

We further make an assumption similar to Assumption 4.1 for the general case:

Assumption 4.4 (Separable distributions)**.** A set of agents E satisfy separable distributions if they can be partitioned into M subsets $S_1, \ldots, S_M$ with $w_1^*, ..., w_M^*$ representing the center of each set respectively, and the optimal parameter θ^* of each local loss $\mathcal{L}_e$ (i.e., $\theta_e^* = \arg\min_\theta \mathcal{L}_e(\theta)$) satisfy $\|\theta_e^* - w_m^*\|_2 \leq r$. In the meantime, agents from different subsets have different data distributions, such that $\|w_{m_1}^* - w_{m_2}^*\|_2 \geq R, \forall m_1, m_2 \in [M], m_1 \neq m_2$.

Theorem 4.5. *Assume the agent set E satisfies the separable distributions condition in Assumption 4.4. Suppose loss functions have bounded variance for gradients on local datasets, i.e.,* $\mathbb{E}_{(x,y)\sim\mathcal{D}_e}[\|\nabla \ell(x, y; \theta) - \nabla \mathcal{L}_e(\theta)\|_2^2] \leq \sigma^2$*, and the population losses are bounded, i.e.,* $\mathcal{L}_e \leq G, \forall e \in [E]$*. With* $\pi_{em}^{(0)} = \frac{1}{M}$, $\exists \Delta_0 > 0, \|w_m^{(0)} - w_m^*\|_2 \leq \frac{\sqrt{\mu}R}{\sqrt{\mu}+\sqrt{L}} - r - \Delta_0$*, and the learning rate of each agent* $\eta \leq \min(\frac{1}{2(\mu+L)}, \frac{\beta}{\sqrt{T}})$*,FOCUS converges by*

$$\pi_{em}^{(T)} \geq \frac{1}{1 + (M-1)\exp(-\mu R \Delta_0 T)}, \ \forall e \in S_m, \tag{9}$$

$$\mathbb{E}\|w_m^{(T)} - w_m^*\|_2^2 \leq (1 - \eta A)^{KT}(\|w_m^{(0)} - w_m^*\|_2^2 + B) + O(Kr) + ME\beta O(K^3, \frac{\sigma^2}{n_e})T^{-1/2} \tag{10}$$

where T is the total number of communication rounds; K is the number of local updates in each communication round; $\gamma_m = |S_m|$ is the number of agents in the m-th cluster, and

$$\underbrace{A = \frac{2\gamma_m}{M}\frac{\mu L}{\mu + L}}_{\textit{related to convergence rate}}, \underbrace{B = \frac{GMTE(\frac{4L}{\mu} + \frac{6}{\mu(\mu+L)})}{(1 - \eta A)^K - \exp(-\mu R \Delta_0)}}_{\textit{caused by the offset of initial clustering}}. \tag{11}$$

Remarks. Theorem 4.5 extends the convergence guarantee of (Π, W) from linear models (Theorem 4.2) to general models with smooth and convex loss functions. For any agent e that in cluster m ($e \in S_m$), its soft cluster label π_{em} converges to 1 based on Equation (69), indicating the clustering optimality. Its model weights W converge linearly to a near-optimal solution. The error term $O(Kr)$ in Equation (70) is expected since r represents the data divergence within each cluster and w_m^* denotes the center of each cluster. The last term in Equation (70) implies a trade-off between communication cost and convergence speed. Increasing K reduces communication cost by $O(\frac{1}{K})$ but at the expanse of slowing down the convergence. Detailed proofs are deferred to Appendix A.2.3.

4.2 Fairness Analysis

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

Linear models. We first concretize such a scenario for linear models. Suppose we have E agents learning weights for M linear models. Their local data $D_e (e \in [E])$ 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 $\epsilon_e^{(i)} \sim \mathcal{N}(0, \sigma_e^2)$. $E - 1$ agents learn from a normal dataset with ground truth vector $\mu_1, \ldots, \mu_{E-1}$ and $\|\mu_e - \mu^*\|_2 \leq r$, 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 stated in Theorem 4.2, the soft clustering labels and 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 the FAA achieved by FOCUS and FedAvg to underscore how our algorithm helps improve fairness for heterogeneous agents.

Theorem 4.6. *When a single agent has an outlier distribution, the fairness FAA achieved by FOCUS algorithm with two clusters $M = 2$ is*

$$\mathcal{FAA}_{focus}(W, \Pi) \leq \delta^2 r^2. \tag{12}$$

while the fairness FAA achieved by FedAvg is

$$\mathcal{FAA}_{avg}(W) \geq \delta^2 \Big(\frac{R^2(E-2) - 2Rr}{E} + r^2\Big) = \Omega(\delta^2 R^2). \tag{13}$$

Remarks. The fairness gap between Fedavg and FOCUS with a single outlier is

$$\mathcal{FAA}_{avg}(W) - \mathcal{FAA}_{focus}(W,\Pi) \geq \delta^2\Big(\frac{R^2(E-2)-2Rr}{E}\Big). \tag{14}$$

As long as $R > \frac{2r}{E-2}$, FOCUS is guaranteed to achieve stronger fairness (i.e., lower FAA) than FedAvg. Note that *the outlier assumption only makes sense when* $E > 2$ since one cannot tell which agent is the outlier when $E = 2$. Also, we naturally assume $R > 2r$ so that the two underlying clusters are at least separable. Thus, we conclude that FOCUS dominates than FedAvg in terms of FAA. Here we discuss a single outlier agent scenario for clarity, and similar conclusions hold for multiple underlying clusters and $M > 2$, as shown in Appendix B.1.

Smooth and strongly convex loss functions. We generalize the fairness analysis to nonlinear models with smooth and convex loss functions. 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 4.3) and bounded, i.e., $\mathcal{L}_e(\theta) \leq G$. $E-1$ agents learn from similar data distributions, such that the total variation distance between the distributions of any two different agents $i, j \in [E-1]$ is no 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 the Bayes error $\mathcal{L}_E(\theta_i^*) - \mathcal{L}_E(\theta_E^*) \geq R$ for any $i \in [E-1]$. We claim that this assumption can be reduced to a lower bound on H-divergence [45] 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 4.7. *The fairness FAA achieved by FOCUS with two clusters* $M = 2$ *is*

$$\mathcal{FAA}_{focus}(W,\Pi) \leq \frac{2Gr}{E-1} \tag{15}$$

Let $B = \frac{2Gr}{E-1}$. *The fairness achieved by FedAvg is*

$$\mathcal{FAA}_{avg}(W) \geq \Big(\frac{E-1}{E} - \frac{L}{\mu E^2}\Big)R - \Big(1 + \frac{L(E-1)}{\mu E} - \frac{L^2}{\mu^2 E}\Big)B - \frac{2L}{\mu E}\sqrt{B(R - \frac{L}{\mu}B)} \tag{16}$$

Remarks. Notably, when the outlier distribution is very different from the normal distribution, such that $R \gg Gr$ (which means $B \ll R$), we simplify Equation (134) as

$$\mathcal{FAA}_{avg}(W) \geq \left(\frac{E-1}{E} - \frac{L}{\mu E^2}\right)R. \tag{17}$$

Note that $\mathcal{FAA}_{focus}(W,\Pi) \leq B \ll R$, so the fairness FAA achieved by FedAvg is always larger (weaker) than that of FOCUS, as long as $E \geq \sqrt{L/\mu}$, indicating the effectiveness of FOCUS.

5 Experimental Evaluation

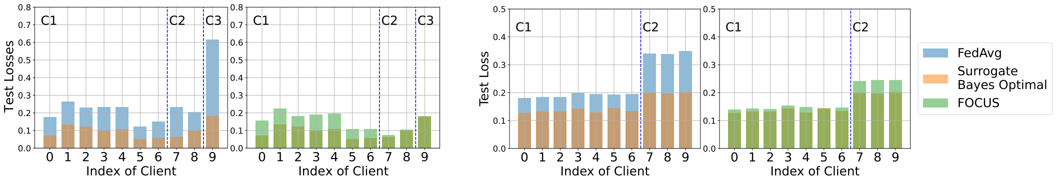
5.1 Experimental Setup

Data and Models. We consider four different datasets with heterogeneous data settings, ranging from synthetic data for linear models to images (rotated MNIST [8] and rotated CIFAR [19]) to text data for sentiment classification on Yelp [43] and IMDb [25] datasets. We train an MLP model for MNIST, a ResNet 18 model [16] for CIFAR, and a pre-trained BERT-base model [9] for the text data. We refer the readers to Appendix D.1 for more implementation details.

Evaluation Metrics and Baselines. We consider following evaluation metrics: average test accuracy, average test loss, FAA for fairness, and the existing fairness metric "agnostic loss" introduced by [29] and "accuracy parity" introduced by [22]. We compare FOCUS to FedAvg and state-of-the-art fair FL algorithms (i.e., q-FFL [22], AFL [29], Ditto [23], and CGSV [38]). We defer the comparision to other FL algorithms under heterogeneous data settings [21, 35, 40] in Appendix D.3. To evaluate FAA of different algorithms, we estimate the Bayes optimal loss $\min_w \mathcal{L}_e(w)$ for each local agent e. Specifically, we train a centralized model based on the 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 agent pair with the maximal difference of excess risks to measure FAA fairness.

Table 1: Comparison of FOCUS, FedAvg, and fair FL algorithms q-FFL, AFL, Ditto and CGSV, in terms of average test accuracy (Avg Acc), average test loss (Avg Loss), fairness FAA and existing fairness metric Agnostic Loss and Accuracy Parity. FOCUS achieves the best fairness measured by FAA compared with all baselines.

		FOCUS	FedAvg	q-FFL			AFL	Ditto	CGSV
				$q = 0.1$	$q = 1$	$q = 10$			
Synthetic	Avg Loss	**0.010**	0.108	0.106	0.102	0.110	0.104	0.023	0.260
	FAA	**0.001**	0.958	0.769	0.717	0.699	0.780	0.012	0.010
	Loss Parity	**4e-5**	0.295	0.261	0.238	0.235	0.244	0.004	0.003
Rotated MNIST	Avg Acc	**0.953**	0.929	0.922	0.861	0.685	0.885	0.940	0.938
	Avg Loss	**0.152**	0.246	0.269	0.489	1.084	0.429	0.210	0.222
	FAA	**0.094**	0.363	0.388	0.612	0.253	0.220	0.104	0.210
	Agnostic Loss	**0.224**	0.616	0.656	1.018	1.271	0.548	0.354	0.331
	Accuracy Parity	**0.014**	0.049	0.052	0.074	0.057	0.032	0.020	0.023
Rotated CIFAR	Avg Acc	**0.688**	0.654	0.648	0.592	0.121	0.661	0.657	0.515
	Avg Loss	**1.133**	2.386	1.138	1.141	2.526	1.666	2.382	3.841
	FAA	**0.360**	1.115	0.620	0.473	0.379	0.595	0.758	1.317
	Agnostic Loss	**1.294**	3.275	1.610	1.439	2.526	2.179	3.053	3.841
	Accuracy Parity	0.027	0.049	0.074	0.069	**0.009**	0.061	0.040	0.022
Yelp/IMDb	Avg Acc	**0.940**	**0.940**	0.938	0.938	0.909	0.934	0.933	0.701
	Avg Loss	**0.174**	0.236	0.188	0.179	0.264	0.187	0.191	0.547
	FAA	**0.047**	0.098	0.052	0.051	0.070	0.049	0.049	0.462
	Agnostic Loss	0.257	0.349	0.266	0.253	**0.242**	0.253	0.263	0.700
	Accuracy Parity	0.015	0.018	0.017	0.016	**0.005**	0.019	0.021	0.171



(a) MNIST

(b) Yelp/IMDb

Figure 1: The excess risks of different agents trained with FedAvg and FOCUS on MNIST (a) and Yelp/IMDb text data (b). C_i denotes ith cluster.

5.2 Evaluation Results

Synthetic data for linear models. We first evaluate FOCUS on linear regression models with synthetic datasets. We fix $E = 10$ agents with data sampled from Gaussian distributions. We study the case considered in Section 4.2 where a single agent has an outlier data distribution, and set the intra-cluster distance $r = 0.01$ and the inter-cluster distance $R = 1$. Table 1 shows that FOCUS achieves FAA of 0.001, much lower than the 0.958 achieved by FedAvg and 0.699 by q-FFL.

Rotated MNIST and CIFAR. Following [13], 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, two subsets are rotated for 90 degrees, one subset is rotated for 180 degrees, and the rest seven subsets are unchanged, yielding an FL setup with three ground-truth clusters. Similarly, for CIFAR, we fix the images of 7 subsets and rotate the other 3 subsets for 180 degrees, thus creating two ground-truth clusters. From Table 1, we observe that FOCUS consistently achieves higher average test accuracy, lower average test loss, and lower FAA than other methods on both datasets. In addition, although existing fair algorithms q-FFL and AFL achieve lower FAA scores than FedAvg, their average test accuracy drops significantly. This is mainly because these fair algorithms are designed for performance parity via improving low-quality agents (i.e., agents with high training loss), thus sacrificing the accuracy of high-quality agents. In contrast, FOCUS improves both the FAA fairness and preserves high test accuracy.

We analyze the surrogate excess risk of each agent on MNIST in Figure 1 (a). The global model trained by FedAvg has the highest test loss of 0.61 on the outlier cluster (C3), resulting in high excess risk for the 9th agent. The low-quality data of the outlier cluster affect the agents in the 1st cluster via FedAvg training, resulting in much higher excess risk than that of FOCUS. FOCUS successfully identifies outlier clusters (2 and 3), rendering models trained from them independent from normal cluster 1. As shown in Figure 1, FOCUS reduces excess risks of all agents, especially the outliers, on different datasets, leading to strong fairness in terms of FAA. 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 D.10.

Sentiment classification. We evaluate FOCUS on the sentiment classification task with text data, Yelp (restaurant reviews), and IMDb (movie reviews), which naturally form data heterogeneity among 10 agents and thus create 2 clusters. Specifically, we sample 56k reviews from Yelp datasets distributed among seven agents and use the whole 25k IMDB datasets distributed among three agents to simulate the heterogeneous setting. From Table 1, we can see that while the average test accuracy of FOCUS, FedAvg, and other fair FL algorithms are similar, FOCUS achieves a lower average test loss. In addition, the FAA of FOCUS is significantly lower than other baselines, indicating stronger fairness. We also observe from Figure 1 (b) that the excess risk of FOCUS on the outlier cluster (i.e., C2) drops significantly compared with that of FedAvg.

Ablation Studies. To provide a more comprehensive evaluation for FOCUS, we present additional ablation studies on scalability (Appendix D.4), performance on different number of outliers (Appendix D.5), convergence rate (Appendix D.6), and runtime analysis (Appendix D.7). The results show that FOCUS is scalable to larger client groups and consumes comparable running time with other methods on different datasets. Moreover, we compare FOCUS to a cluster-wise FedAvg algorithm with hard clustering, illustrating the advantages of FOCUS using soft-clustering when the underlying clusters are not perfectly separable. We refer readers to Appendix D.8 for further discussions. In addition, we analyze the effect of choosing M in practice Appendix D.9. We evaluate FOCUS on the FEMNIST dataset [6], where data distributions exhibit ambiguous cluster structure, showing the robustness of FOCUS against the underlying cluster structures.

6 Conclusion

In this work, we provide an agent-level fairness measurement in FL (FAA) by taking agents' inherent heterogeneous data properties into account. Motivated by our fairness definition in FL, we provide an effective FL training algorithm FOCUS to achieve high fairness. We provide theoretical analysis for the convergence and fairness of FOCUS, and empirically show that FOCUS achieves stronger fairness than existing FL methods, while achieving similar or higher prediction accuracy.

References

[1] Mohammed Adnan, Shivam Kalra, Jesse Cresswell, Graham Taylor, and Hamid Tizhoosh. Federated learning and differential privacy for medical image analysis. volume 12, 02 2022. doi: 10.1038/s41598-022-05539-7.

[2] 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. URL `http://www.diva-portal.org/smash/get/diva2:1584574/FULLTEXT01.pdf`.

[3] 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. URL `https://sisinflab.poliba.it/publications/2021/ADDFN21b/paper7.pdf`.

[4] Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[6] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

[7] Lingyang Chu, Lanjun Wang, Yanjie Dong, Jian Pei, Zirui Zhou, and Yong Zhang. Fedfair: Training fair models in cross-silo federated learning. 2021. URL https://arxiv.org/pdf/2109.05662.pdf.

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

[9] 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.

[10] Kate Donahue and Jon Kleinberg. Models of fairness in federated learning. 2022. URL https://arxiv.org/pdf/2112.00818.pdf.

[11] Kate Donahue and Jon Kleinberg. Models of fairness in federated learning. 2022. URL https://arxiv.org/abs/2112.00818.

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[13] 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.

[14] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

[15] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[16] 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.

[17] Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. Provably fair federated learning via bounded group loss. 2022. URL https://arxiv.org/pdf/2203.10190.pdf.

[18] 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. URL https://arxiv.org/pdf/1905.07479.pdf.

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

[20] Chengxi Li, Gang Li, and Pramod K. Varshney. Federated learning with soft clustering. 2022. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9174890.

[21] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

[22] 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. URL https://openreview.net/forum?id=ByexElSYDr.

[23] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pages 6357–6368. PMLR, 2021.

[24] Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 22(3):2031–2063, 2020.

[25] 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. URL `http://www.aclweb.org/anthology/P11-1015`.

[26] 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. URL `https://openreview.net/forum?id=YCqx6zhEzRp`.

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

[28] 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. URL `https://doi.org/10.1145/3460231.3474262`.

[29] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. 2019. URL `http://proceedings.mlr.press/v97/mohri19a/mohri19a.pdf`.

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

[31] Yichen Ruan and Carlee Joe-Wong. Fedsoft: Soft clustered federated learning with proximal local updating. 2022. URL `https://arxiv.org/pdf/2112.06053.pdf`.

[32] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. 2021. URL `https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9174890`.

[33] 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. URL `https://doi.org/10.1038/s41598-020-69250-1`.

[34] Morris Stallmann and Anna Wilbik. Towards federated clustering: A federated fuzzy c-means algorithm (ffcm). 2022. URL `https://arxiv.org/pdf/2201.07316.pdf`.

[35] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*, 2020.

[36] Zheng Wang, Xiaoliang Fan, Jianzhong Qi, Chenglu Wen, Cheng Wang, and Rongshan Yu. Federated learning with fair averaging. 2021. URL `https://www.ijcai.org/proceedings/2021/0223.pdf`.

[37] Ming Xie, Guodong Long, Tao Shen, Tianyi Zhou, Xianzhi Wang, Jing Jiang, and Chengqi Zhang. Multi-center federated learning. 2021. URL `https://arxiv.org/abs/2005.01026`.

[38] Xinyi Xu, Lingjuan Lyu, Xingjun Ma, Chenglin Miao, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Gradient driven rewards to guarantee fairness in collaborative machine learning. *Advances in Neural Information Processing Systems*, 34:16104–16117, 2021.

[39] Bowei Yan, Mingzhang Yin, and Purnamrita Sarkar. Convergence of gradient em on multi-component mixture of gaussians. *Advances in Neural Information Processing Systems*, 30, 2017.

[40] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*, pages 7252–7261. PMLR, 2019.

[41] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[42] Jingfeng Zhang, Cheng Li, Antonio Robles-Kelly, and Mohan Kankanhalli. Hierarchically fair federated learning. 2020. URL `https://arxiv.org/pdf/2004.10386.pdf`.

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

[44] Han Zhao and Geoffrey J. Gordon. Inherent tradeoffs in learning fair representations. *Journal of Machine Learning Research*, 23(57):1–26, 2022. URL `http://jmlr.org/papers/v23/21-1427.html`.

[45] 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. URL `https://openreview.net/forum?id=KB5onONJIAU`.

A Convergence Proof

A.1 Convergence of Linear Models (Theorem 4.2)

A.1.1 Key Lemmas

We need to state two lemmas first before proving Theorem 4.2.

Lemma A.1. *Suppose $e \in S_m$ and the m-th cluster is the one closest to w_m^*. Assume $\|w_m^{(t)} - w_m^*\| \leq \alpha < \beta \leq \min_{m' \neq m} \|w_{m'}^{(t)} - w_m^*\|$. Then the E-step updates as*

$$\pi_{em}^{(t+1)} \geq \frac{\pi_{em}^{(t)}}{\pi_{em}^{(t)} + (1 - \pi_{em}^{(t)}) \exp\Big(- (\beta^2 - \alpha^2 - 2(\alpha + \beta) r) \delta^2 \Big)} \tag{18}$$

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

$$\pi_{em}^{(t+1)} \geq \frac{\pi_{em}^{(t)}}{\pi_{em}^{(t)} + (1 - \pi_{em}^{(t)}) \exp(-2R\Delta\delta^2)}, \quad \forall e \in S_m \tag{19}$$

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

$$\theta_{em}^{(k+1)} = \theta_{em} - \eta_k g_{em}(\theta_{em}) = \theta_{em}^{(k)} - \eta_k \nabla \sum_{i=1}^{n_e} \ell(h_{\theta_{em}}(x_e^{(i)}), y_e^{(i)}). \tag{20}$$

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

$$\hat{w}_m^{(k)} = \sum_{e=1}^{E} \frac{\pi_{em} \theta_{em}^{(k)}}{\sum_{e'=1}^{E} \pi_{e'm}}. \tag{21}$$

Lemma A.2. *Suppose any agent $e \in S_m$ has a soft clustering label $\pi_{em}^{(t+1)} \geq p$. Then one step of local SGD updates $\hat{w}_m^{(k)}$ by Equation (22), if the learning rate $\eta_k \leq \frac{1}{4\delta^2}$.*

$$\mathbb{E}\|\hat{w}_m^{(k+1)} - w_m^*\|_2^2 \leq (1 - 2\eta_k \gamma_m p \delta^2) \mathbb{E}\|\hat{w}_m^{(k+1)} - w_m^*\|_2^2 + \eta_k A_1 + \eta_k^2 A_2. \tag{22}$$

$$A_1 = 4\gamma_m r \delta^2 + 2\delta^2 E(1-p), A_2 = 16E(K-1)^2\delta^4 + O(\frac{d}{n_e}) E(\delta^4 + \delta^2\sigma^2) \tag{23}$$

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

$$\mathbb{E}\|\hat{w}_m^{(k)} - w_m^*\|_2^2 \leq (1 - 2\eta\gamma_m p \delta^2)^k \mathbb{E}\|\hat{w}_m^{(0)} - w_m^*\|_2^2 + \eta k (A_1 + \eta A_2). \tag{24}$$

A.1.2 Completing the Proof of Theorem 4.2

We now combine Lemma A.1 and Lemma A.2 to prove Theorem 4.2. The theorem is restated below.

Theorem 4.2. *With the assumptions 1 and 2, $n_e = O(d)$, if learning rate $\eta \leq \min(\frac{1}{4\delta^2}, \frac{\beta}{\sqrt{T}})$,*

$$\pi_{em}^{(T)} \geq \frac{1}{1 + (M-1) \cdot \exp(-2R\delta^2\Delta_0 K)}, \forall e \in S_m \tag{25}$$

$$\mathbb{E}\|w_m^{(T)} - w_m^*\|_2^2 \leq (1 - \frac{2\eta\gamma_m\delta^2}{M})^{KT} (\|w_m^{(0)} - w_m^*\|_2^2 + A) + 2MKr + \frac{M\delta^2 E\beta}{2\sqrt{T}} O(K^3, \sigma^2). \tag{26}$$

where K is the total number of communication rounds; T is the number of iterations each round; $\gamma_m = |S_m|$ is the number of agents in the m-th cluster, and

$$A = \frac{2EK(M-1)\delta^2}{(1 - \frac{2\eta\delta^2\gamma_m}{M})^K - \exp(-2R\delta^2\Delta_0)}. \tag{27}$$

Proof. We prove Theorem 4.2 by induction. Suppose

$$
\begin{aligned}
&\pi_{em}^{(t)} \geq \frac{1}{1+(M-1)\exp(-2R\delta^2\Delta_0 t)} && (28)\\
&\mathbb{E}\|w_m^{(t)} - w_m^*\|^2 \leq (1-\frac{2\eta\gamma_m\delta^2}{M})^{Kt}(\|w_m^{(0)} - w_m^*\|^2) + A\Big((1-\frac{2\eta\gamma_m\delta^2}{M})^{Kt} - \exp\big(-2R\delta^2\Delta_0 t\big)\Big)\\
&\qquad\qquad + \frac{\eta B}{1-(1-\frac{2\eta\gamma_m\delta^2}{M})^K}. && (29)
\end{aligned}
$$

where $B = [16E\delta^4K^3 + EK(\delta^4 + \delta^2\sigma^2)]\eta + 4\gamma_m r\delta^2 K$.

Then according to Lemma A.1,

$$
\begin{aligned}
\pi_{em}^{(t+1)} &\geq \frac{\pi_{em}^{(t)}}{\pi_{em}^{(t)} + (1-\pi_{em}^{(t)})\exp(-2R\Delta_0\delta^2)} && (30)\\
&\geq \frac{1}{1+(M-1)\exp(-2R\delta^2\Delta_0 t)\exp(-2R\Delta_0\delta^2)} && (31)\\
&\geq \frac{1}{1+(M-1)\exp(-2R\Delta_0\delta^2(t+1))}. && (32)
\end{aligned}
$$

We recall the virtual sequence of $\hat{w}_m$ defined by Equation (21). Since models are synchronized after K rounds, the know $\hat{w}_m^{(0)} = w_m^{(t)}$ and $w_m^{(t+1)} = \hat{w}_m^{(K)}$. We then apply Lemma A.2 to prove the induction. Note that instead of proving Equation (26), we prove a stronger induction hypothesis of Equation (29).

$$
\begin{aligned}
&\mathbb{E}\|w_m^{(t+1)} - w_m^*\|^2\\
&= \mathbb{E}\|\hat{w}_m^{(K)} - w_m^*\|^2 && (33)\\
&\leq (1-2\eta\gamma_m p\delta^2)^K\mathbb{E}\|\hat{w}_m^{(t)} - w_m^*\|^2 + \eta K(A_1 + \eta A_2) && (34)\\
&\leq (1-2\eta\gamma_m p\delta^2)^K\Big((1-\frac{2\eta\gamma_m\delta^2}{M})^{Kt}\|w_m^{(0)} - w_m^*\|^2 + A((1-\frac{2\eta\gamma_m\delta^2}{M})^{Kt} - \exp\big(-2R\Delta_0\delta^2 t\big))\\
&\qquad + \frac{\eta B}{1-(1-\frac{2\eta\gamma_m\delta^2}{M})^K}\Big) + \eta K(4\gamma_m r\delta^2 + 2\delta^2E(1-p)) + \eta^2KA_2 && (35)\\
&\leq (1-\frac{2\eta\gamma_m\delta^2}{M})^{(t+1)K}\|w_m^{(0)} - w_m^*\|^2\\
&+ \underbrace{A(1-\frac{2\eta\gamma_m\delta^2}{M})^{(t+1)K} - A\exp\big(-2R\Delta_0\delta^2 t\big)(1-\frac{2\eta\gamma_m\delta^2}{M})^K + 2\delta^2E(1-p)}_{D_1}\\
&+ \underbrace{(1-\frac{2\eta\gamma_m\delta^2}{M})^K\frac{\eta B}{1-(1-\frac{2\eta\gamma_m\delta^2}{M})^K} + 4\eta K\gamma_m r\delta^2 + \eta^2KA_2}_{D_2}. && (36)
\end{aligned}
$$

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

$$
\begin{aligned}
D_1 &\leq A(1-\frac{2\eta\gamma_m\delta^2}{M})^{(t+1)K} - A\exp\big(-2R\Delta_0\delta^2 t\big)(1-\frac{2\eta\gamma_m\delta^2}{M})^K + 2\delta^2EK(M-1)\exp\big(-2R\Delta_0\delta^2 t\big)\\
&\leq A((1-\frac{2\eta\gamma_m\delta^2}{M})^{(t+1)K} - \exp\big(-2R\Delta_0\delta^2(t+1)\big)) && (37)
\end{aligned}
$$

For D_2 we have

$$
\begin{aligned}
D_2 &\leq (1-\frac{2\eta\gamma_m\delta^2}{M})^K\frac{\eta B}{[1-(1-\frac{2\eta\gamma_m\delta^2}{M})^K]} + 4\eta\gamma_m r\delta^2K + 16\eta^2E\delta^4K^3 + \eta^2EKO(\delta^4+\delta^2\sigma^2)\\
&= \frac{\eta B}{1-(1-\frac{2\eta\gamma_m\delta^2}{M})^K}. && (38)
\end{aligned}
$$

Finally we combine Equations (36) to (38) so

$$\mathbb{E}\|w_m^{(t+1)} - w_m^*\|^2 \leq (1 - \frac{2\eta\gamma_m\delta^2}{M})^{(t+1)K}\|w_m^{(0)} - w_m^*\|^2 + A\Big((1 - \frac{2\eta\gamma_m\delta^2}{M})^{(t+1)K} - \exp\big(-2R\delta^2\Delta_0(t+1)\big)\Big) + \frac{\eta B}{1-(1-\frac{2\eta\gamma_m\delta^2}{M})^K}. \tag{39}$$

Since it is trivial to check that both induction hypotheses hold when $t = 0$, the induction hypothesis holds. Note that $K \geq 1$, so

$$\frac{\eta B}{1-(1-\frac{2\eta\gamma_m\delta^2}{M})^K} \leq \eta B\frac{M}{2\eta\gamma_m\delta^2} \leq 2MKr + \frac{M\delta^2 E\beta}{2\sqrt{T}}O(K^3, \delta^2). \tag{40}$$

Combining Equation (39) and Equation (40) completes our proof. □

A.1.3 Deferred Proofs of Key Lemmas

Lemma 1.

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

$$\pi_{em}^{(t+1)} = \frac{\pi_{em}^{(t)}\exp\big[-\mathbb{E}_{(x,y)\sim D_e}({w_m}^Tx - y)^2\big]}{\sum_{m'}\pi_{em'}^{(t)}\exp\big[-\mathbb{E}_{(x,y)\sim D_e}({w_{m'}}^Tx - y)^2\big]} \tag{41}$$

so

$$\begin{aligned}
\pi_{em}^{(t+1)} &= \frac{\pi_{em}^{(t)}\exp\Big(-\|w_m^{(t)} - \mu_e\|^2\delta^2\Big)}{\sum_{m'}\pi_{em'}^{(t)}\exp\Big[-\|{w'_m}^{(t)} - \mu_e\|^2\delta^2\Big]} &(42)\\
&\geq \frac{\pi_{em}^{(t)}\exp\big(-(\beta - r)^2\delta^2\big)}{\pi_{em}^{(t)}\exp(-(\beta-r)^2\delta^2) + \sum_{m'\neq m}\pi_{em'}^{(t)}\exp(-(\alpha+r)^2\delta^2)} &(43)\\
&\geq \frac{\pi_{em}^{(t)}}{\pi_{em}^{(t)} + (1-\pi_{em}^{(t)})\exp\Big(-(\beta^2 - \alpha^2 - 2(\alpha+\beta)r)\delta^2\Big)} &(44)
\end{aligned}$$

□

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,

$$\begin{aligned}
\|\hat{w}_m^{(k+1)} - w_m^*\|^2 &= \|w_m^{(k)} - w_m^* - \eta_k g_k\|^2 &(45)\\
&= \|\hat{w}_m^{(k)} - w_m^* - \eta_k\frac{2}{n_e}\sum_e\pi_{em}X_e^TX_e(\theta_{em}^{(k)} - \mu_e) + \frac{2\eta_k}{n_e}\sum_e\pi_{em}X_e^T\epsilon_e\|^2 &(46)\\
&= \|\hat{w}_m^{(k)} - w_m^* - \hat{g}_k\|^2 + \eta_k^2\|g_k - \hat{g}_k\|^2 + 2\eta_k\langle w_m^{(k)} - w_m^* - \hat{g}_k, \hat{g}_k - g_k\rangle. &(47)
\end{aligned}$$

where $\hat{g}_k = \frac{2}{n_e}\sum_e\pi_{em}\mathbb{E}(X_e^TX_e)(\theta_{em}^{(k)} - \mu)$. Since the expectation of the last term in Equation (47) is zero, we only need to estimate the expectation of $\|\hat{w}_m^{(k)} - w_m^* - \eta_k\hat{g}_k\|^2$ and $\|\hat{g}_k - g_k\|^2$.

$$
\begin{aligned}
&\|\hat{w}_m^{(k)} - w_m^* - \eta_k \hat{g}_k\|^2 \\
&= \|\hat{w}_m^{(k)} - w_m^*\|^2 + \frac{4\eta_k^2}{n_e^2} \sum_e \pi_{em} \mathbb{E}(X_e^T X_e) \|\theta_{em}^t - \mu_e\|^2 - \frac{4\eta_k}{n_e} \sum_e \pi_{em} \langle \hat{w}_m^{(k)} - w_m^*, \mathbb{E}(X_e^T X_e)(\theta_{em}^{(k)} - \mu_e) \rangle \\
&= \|\hat{w}_m^{(k)} - w_m^*\|^2 + 4\eta_k^2 \delta^2 \sum_e \pi_{em} \|\theta_{em}^{(k)} - \mu_e\|^2 - \underbrace{4\eta_k \langle \hat{w}_m^{(k)} - w_m^*, \sum_e \pi_{em} \delta^2 (\theta_{em}^{(k)} - \mu_e) \rangle}_{C_1} .
\end{aligned} \tag{48}
$$

$$
C_1 = -4\eta_k \sum_e \pi_{em} \langle \hat{w}_m^{(k)} - \theta_{em}^{(k)}, \delta^2 (\theta_{em}^{(k)} - \mu_e) \rangle - 4\eta_k \sum_e \pi_{em} \langle \theta_{em}^{(k)} - w_m^*, \delta^2 (\theta_{em}^{(k)} - \mu_e) \rangle \tag{49}
$$

$$
\begin{aligned}
&\le 4 \sum_e \pi_{em} \|\hat{w}_m^{(k)} - \theta_{em}^{(k)}\|^2 + 4\delta^4 \eta_k^2 \sum_e \pi_{em} \|\theta_{em}^{(k)} - \mu_e\|^2 - 4\eta_k \delta^2 \sum_e \pi_{em} \|\theta_{em}^{(k)} - \mu_e\|^2 \\
&\qquad - \underbrace{4\eta_k \delta^2 \sum_e \pi_{em} \langle \mu_e - w_m^*, \theta_{em}^{(k)} - \mu_e \rangle}_{C_2}
\end{aligned} \tag{50}
$$

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

$$
\mathbb{E}\|\hat{w}_m^{(k)} - w_m^* - \eta_k \hat{g}_k\|^2 \tag{51}
$$

$$
\le \mathbb{E}\|\hat{w}_m^{(k)} - w_m^*\|^2 + (8\delta^4 \eta_k^2 - 4\eta_k \delta^2) \sum_e \pi_{em} \mathbb{E}\|\theta_{em}^{(k)} - \mu_e\|^2 + 4 \sum_e \pi_{em} \mathbb{E}\|\hat{w}_m^{(k)} - \theta_{em}^{(k)}\|^2 + C_2 \tag{52}
$$

$$
\le \mathbb{E}\|\hat{w}_m^{(k)} - w_m^*\|^2 - 2\eta_k \delta^2 \sum_e \pi_{em} \mathbb{E}\|\theta_{em}^{(k)} - \mu_e\|^2 + 4 \sum_e \pi_{em} \mathbb{E}\|\hat{w}_m^{(k)} - \theta_{em}^{(k)}\|^2 + C_2 \tag{53}
$$

Note that

$$
\sum_e \pi_{em} \mathbb{E}\|\theta_{em}^{(k)} - \mu_e\|^2 \tag{54}
$$

$$
= \sum_{e \in S_m} \pi_{em} \mathbb{E}\|\theta_{em}^{(k)} - \mu_e\|^2 + \sum_{e \notin S_m} \pi_{em} \mathbb{E}\|\theta_{em}^{(k)} - \mu_e\|^2 \tag{55}
$$

$$
\ge \sum_{e \in S_m} \pi_{em} (\mathbb{E}\|\theta_{em}^{(k)} - w_m^*\|^2 + 2r + r^2) + \sum_{e \notin S_m} \pi_{em} \mathbb{E}\|\theta_{em}^{(k)} - \mu_e\|^2 \tag{56}
$$

$$
= \sum_{e \in S_m} \pi_{em} (\mathbb{E}\|\hat{w}_m^{(k)} - w_m^*\|^2 + \mathbb{E}\|\hat{w}_m^{(k)} - \theta_{em}^{(k)}\|^2 + 2r + r^2) + \sum_{e \notin S_m} \pi_{em} \mathbb{E}\|\theta_{em}^{(k)} - \mu_e\|^2 \tag{57}
$$

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

$$
4\mathbb{E} \sum_e \pi_{em} \|\hat{w}_m^{(k)} - \theta_{em}^{(k)}\|^2 \le 4\mathbb{E} \sum_e \pi_{em} \|\hat{w}_m^{(0)} - \theta_{em}^{(k)}\|^2 \tag{58}
$$

$$
\le 4 \sum_e \pi_{em} (K-1) \mathbb{E} \sum_{t'}^{t-1} {\eta_k'}^2 \|\frac{2}{n_e} X_e^T X_e (\theta_{em}^{(k)} - \mu_e)\|^2 \tag{59}
$$

$$
\le 16\eta_k^2 E(K-1)^2 \delta^4. \tag{60}
$$

Thus,

$$
\begin{aligned}
\mathbb{E}\|\hat{w}_m^{(k)} - w_m^* - \eta_k \hat{g}_k\|^2 \le (1 - 2\eta_k \delta^2 \sum_e \pi_{em}) \mathbb{E}\|\hat{w}_m^{(k)} - w_m^*\|^2 + 16\eta_k^2 E(K-1)^2 \delta^4 \\
\underbrace{-2\eta_k \delta^2 \sum_{e \notin S_m} \pi_{em} \mathbb{E}\|\theta_{em}^{(k)} - \mu_e\|^2 - 4\eta_k \delta^2 \sum_e \pi_{em} \langle \theta_{em}^{(k)} - \mu_e, \mu_e - w_m^* \rangle}_{C_3}
\end{aligned} \tag{61}
$$

Since

$$C_3 \le 2\eta_k\delta^2 \sum_{e\notin S_m} \pi_{em}\|\mu_e - w_m^*\|_2^2 - 4\eta_k\delta^2 \sum_{e\in S_m} \pi_{em}\|\theta_{em}^{(k)} - \mu_e\|_2\|\mu_e - w_m^*\|_2 \tag{62}$$

$$\le 2\eta_k\delta^2 E(1-p) + 4\eta_k\delta^2\gamma_m r \tag{63}$$

we have

$$\mathbb{E}\|\hat{w}_m^{(k)} - w_m^* - \eta_k\hat{g}_k\|^2 \le (2\eta_k\delta^2\gamma_m p)\mathbb{E}\|\hat{w}_m^{(k)} - w_m^*\|^2 + 16\eta_k^2 E(K-1)^2\delta^4 + 2\eta_k\delta^2 E(1-p) + 4\eta_k\delta^2\gamma_m r \tag{64}$$

Notice that

$$\mathbb{E}\|\hat{g}_k - g_k\|^2 = \mathbb{E}\sum_e \frac{4}{n_e^2}\pi_{em}\|(X_e^T X_e - \mathbb{E}(X_e^T X_e))(\theta_{em}^{(k)} - \mu_e)\|^2 + \mathbb{E}\sum_e \frac{4}{n_e^2}\sum_e \pi_{em}\|X_e^T\epsilon_e\|^2$$

$$= E\frac{O(dn_e)}{n_e^2}\delta^4 + E\frac{O(dn_e)}{n_e^2}\delta^2\sigma^2 \tag{65}$$

so

$$\mathbb{E}\|\hat{w}_m^{(k+1)} - w_m^*\|_2^2 \le (1 - 2\eta_k\gamma_m p\delta^2)\mathbb{E}\|\hat{w}_m^{(k)} - w_m^*\|_2^2 + \eta_k A_1 + \eta_k^2 A_2 \tag{66}$$

where

$$A_1 = 4\delta^2\gamma_m r + 2\delta^2 E(1-p) \tag{67}$$

and

$$A_2 = 16E(K-1)^2\delta^4 + O(\frac{d}{n_e})E(\delta^4 + \delta^2\sigma^2). \tag{68}$$

□

A.2 Convergence of Models with Smooth and Strongly Convex Losses (Theorem 4.5)

We restate Theorem 4.5 for clarity here.

Theorem A.3. *Assume the agent set E satisfies the separable distributions condition in Assumption 4.4. Suppose loss functions have bounded variance for gradients on local datasets, i.e., $\mathbb{E}_{(x,y)\sim\mathcal{D}_e}[\|\nabla\ell(x,y;\theta) - \nabla\mathcal{L}_e(\theta)\|_2^2] \le \sigma^2$, and the population losses are bounded, i.e., $\mathcal{L}_e \le G, \forall e \in [E]$. With $\pi_{em}^{(0)} = \frac{1}{M}$, $\exists\Delta_0 > 0, \|w_m^{(0)} - w_m^*\|_2 \le \frac{\sqrt{\mu}R}{\sqrt{\mu}+\sqrt{L}} - r - \Delta_0$, and the learning rate of each agent $\eta \le \min(\frac{1}{2(\mu+L)}, \frac{\beta}{\sqrt{T}})$,FOCUS converges by*

$$\pi_{em}^{(T)} \ge \frac{1}{1 + (M-1)\exp(-\mu R\Delta_0 T)}, \ \forall e \in S_m, \tag{69}$$

$$\mathbb{E}\|w_m^{(T)} - w_m^*\|_2^2 \le (1-\eta A)^{KT}(\|w_m^{(0)} - w_m^*\|_2^2 + B) + O(Kr) + ME\beta O(K^3, \frac{\sigma^2}{n_e})T^{-1/2} \tag{70}$$

where T is the total number of communication rounds; K is the number of local updates in each communication round; $\gamma_m = |S_m|$ is the number of agents in the m-th cluster, and

$$\underbrace{A = \frac{2\gamma_m}{M}\frac{\mu L}{\mu + L}}_{\text{related to convergence rate}}, \underbrace{B = \frac{GMTE(\frac{4L}{\mu} + \frac{6}{\mu(\mu+L)})}{(1-\eta A)^K - \exp(-\mu R\Delta_0)}}_{\text{caused by the offset of initial clustering}}. \tag{71}$$

Here we present the detailed proof for Theorem 4.5.

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 A.4. *Suppose the loss function $\mathcal{L}_{P_t}(\theta)$ is L-smooth and μ-strongly convex for any cluster m. If $\|w_m^{(t)} - w_m^*\| \le \frac{\sqrt{\mu}R}{\sqrt{\mu}+\sqrt{L}} - r - \Delta$ for some $\Delta > 0$, then E-step updates as*

$$\pi_{em}^{(t)} \ge \frac{\pi_{em}^{(t)}}{\pi_{em}^{(t)} + (1 - \pi_{em}^{(t)})\exp(-\mu R\Delta)}. \tag{72}$$

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

$$\theta_{em}^{(k+1)} = \theta_{em} - \eta_k g_{em}(\theta_{em}) = \theta_{em}^{(k)} - \eta_k \nabla \sum_{i=1}^{n_e} \ell(h_{\theta_{em}}(x_e^{(i)}), y_e^{(i)}). \tag{73}$$

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

$$\hat{w}_m^{(k)} = \sum_{e=1}^{E} \frac{\pi_{em}\theta_{em}^{(k)}}{\sum_{e'=1}^{E} \pi_{e'm}}. \tag{74}$$

Lemma A.5. *Suppose for any agent $e \in S_m$, its soft clustering label $\pi_{em}^{(t+1)} \geq p$. Then one step local SGD updates $\hat{w}_m^{(k)}$ by Equation (75), if the learning rate $\eta_k \leq \frac{1}{2(\mu+L)}$.*

$$\mathbb{E}\|\hat{w}_m^{(k+1)} - w_m^*\|_2^2 \leq (1 - \eta_k A_0)\mathbb{E}\|\hat{w}_m^{(k)} - w_m^*\|_2^2 + \eta_k A_1 + \eta_k^2 A_2. \tag{75}$$

where

$$A_0 = \frac{2\gamma_m p \mu L}{\mu + L} \tag{76}$$

$$A_1 = 2\gamma_m L r\sqrt{\frac{2G}{\mu}} + \frac{G(1-p)E}{\mu}(4L + \frac{6}{\mu + L}) + O(r^2). \tag{77}$$

$$A_2 = \frac{4E(K-1)^2 G L^2}{\mu} + \frac{E\sigma^2}{n_e}. \tag{78}$$

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

$$\mathbb{E}\|\hat{w}_m^{(k)} - w_m^*\|^2 \leq (1 - \eta A_0)^k \mathbb{E}\|\hat{w}_m^{(0)} - w_m^*\|^2 + \eta k(A_1 + \eta A_2). \tag{79}$$

A.2.2 Completing the Proof of Theorem 4.5

Theorem 4.5. *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 from assumptions 3 and 4, if each agent chooses learning rate $\eta \leq \min(\frac{1}{2(\mu+L)}, \frac{\beta}{\sqrt{T}})$, the weights (Π, W) converges by*

$$\pi_{em}^{(T)} \geq \frac{1}{1 + (M-1)\exp(-\mu R \Delta_0 T)}, \ \forall e \in S_m \tag{80}$$

$$\mathbb{E}\|w_m^{(T)} - w_m^*\|_2^2 \leq (1 - \eta A)^{KT}(\|w_m^{(0)} - w_m^*\|_2^2 + B) + O(Kr) + \frac{ME\beta O(K^3, \frac{\sigma^2}{n_e})}{\sqrt{T}} \tag{81}$$

where T is the total number of communication rounds; K is the number of iterations each round; $\gamma_m = |S_m|$ is the number of agents in the m-th cluster, and

$$A = \frac{2\gamma_m}{M}\frac{\mu L}{\mu + L}, B = \frac{GMTE(\frac{4L}{\mu} + \frac{6}{\mu(\mu+L)})}{(1 - \eta A)^K - \exp(-\mu R \Delta_0)}. \tag{82}$$

Proof. The proof is quite similar to Theorem 1 for linear models: we follow an induction proof using lemmas 3 and 4. Suppose Equation (80) hold for step t. And suppose

$$\mathbb{E}\|w_m^{(t)} - w_m^*\|_2^2 \leq (1-\eta A)^{Kt}(\|w_m^{(0)} - w_m^*\|_2^2) + B((1-\eta A)^{Kt} - \exp(-\mu R\Delta_0 t)) + \frac{\eta C}{1 - (1 - \eta A)^K}. \tag{83}$$

where

$$C = \frac{4\eta E G K^3 L^2}{\mu} + (2\gamma_m L r\sqrt{\frac{2G}{\mu}} + O(r^2)) + \eta\frac{EK\sigma^2}{n_e}. \tag{84}$$

Then for any $e \in S_m$,

$$\pi_{em}^{(t+1)} \geq \frac{\pi_{em}^{(t)}}{\pi_{em}^{(t)} + (1-\pi_{em}^{(t)})\exp(-\mu R \Delta_t)} \tag{85}$$

$$\geq \frac{1}{1+(M-1)\exp(-\mu R\Delta_0 t)\exp(-\mu R \Delta_t)} \tag{86}$$

$$\geq \frac{1}{1+(M-1)\exp(-\mu R\Delta_0 (t+1))} \tag{87}$$

We recall the virtual sequence $\hat{w}_m^{(k)}$ defined in Equation (74). Models are synchronized after K rounds of local iterations, so $w_m^{(t+1)} = \hat{w}_m^{(K)}$. Thus, according to Lemma A.5,

$$\mathbb{E}\|w_m^{(t+1)} - w_m^*\|_2^2 = \mathbb{E}\|\hat{w}_m^{(K)} - w_m^*\|_2^2 \tag{88}$$

$$\leq (1-\eta A_0)^K \mathbb{E}\|w_m^{(t)} - w_m^*\|_2^2 + \eta K(A_1 + \eta A_2) \tag{89}$$

$$\leq (1-\eta A_0)^K \Big((1-\eta A)^{Kt}(\mathbb{E}\|w_m^{(0)} - w_m^*\|^2) + B((1-\eta A)^{Kt} - \exp(-\mu R\Delta_0 t)) + \frac{\eta C}{1-(1-\eta A)^K}\Big) + \eta K(A_1+\eta A_2) \tag{90}$$

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

$$+ \underbrace{(1-\eta A)^K \frac{\eta C}{1-(1-\eta A)^K} + \eta K(2\gamma_m Lr\sqrt{\frac{2G}{\mu}} + O(r^2)) + \eta^2 K A_2}_{F_2}. \tag{91}$$

For F_1, we use the fact that

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

so

$$F_1 \leq (1-\eta A)^K B\big((1-\eta A)^{Kt} - \exp(-\mu R\Delta_0 t)\big) + \eta\frac{G(M-1)\exp(-\mu R\Delta_0 t)}{\mu}(4L + \frac{6}{\mu+L}) \tag{92}$$

$$= B\Big((1-\eta A)^{(t+1)K} - \exp(-\mu R \Delta_0 t)\Big) \tag{93}$$

For F_2, we have

$$F_2 \leq (1-\eta A)^K \frac{\eta C}{1-(1-\eta A)^K} + \eta K(2\gamma_m Lr\sqrt{\frac{2G}{\mu}} + O(r^2)) + \frac{4EGL^2\eta^2K^3}{\mu} + \frac{\eta^2 K E\sigma^2}{n_e} \tag{94}$$

$$\leq \frac{\eta C}{1-(1-\eta A)^K}. \tag{95}$$

Combining F_1 and F_2 finishes the induction proof. Moreover, since $T \geq 1$, we have

$$\frac{\eta C}{1-(1-\eta A)^K} \leq \frac{C}{A} = O(Kr) + \frac{ME\beta}{\sqrt{T}}O(K^3, \frac{\sigma^2}{n_e}). \tag{96}$$

Combining Equation (83) and Equation (96) completes our proof. □

A.2.3 Deferred Proofs of Key Lemmas

Lemma 3.

Proof. According to Algorithm 1,

$$\begin{aligned}
\pi_{em}^{(t+1)} &= \frac{\pi_{em}^{(t)}}{\pi_{em}^{(t)} + \sum_{m' \neq m} \pi_{em'}^{(t)} \exp\Big(\mathbb{E}\ell(x, y; w_m^{(t)}) - \mathbb{E}\ell(x, y; w_{m'}^{(t)})\Big)} && (97)\\
&\geq \frac{\pi_{em}^{(t)}}{\pi_{em}^{(t)} + (1 - \pi_{em}^{(t)}) \exp\Big(\max_{m' \neq m}(\mathcal{L}_{P_e}(w_m^{(t)}) - \mathcal{L}_{P_e}(w_{m'}^{(t)}))\Big)} && (98)
\end{aligned}$$

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

$$\begin{aligned}
\mathcal{L}_{P_e}(w_m^{(t)}) - \mathcal{L}_{P_e}(w_{m'}^{(t)}) &\leq \frac{L}{2}\|w_m^{(t)} - \theta_t^*\|^2 - \frac{\mu}{2}\|w_{m'}^{(t)} - \theta_t^*\|^2\\
&\leq \frac{L}{2}(\frac{\sqrt{\mu}R}{\sqrt{\mu} + \sqrt{L}} - \Delta)^2 - \frac{\mu}{2}(\frac{\sqrt{L}R}{\sqrt{\mu} + \sqrt{L}} + \Delta)^2\\
&\leq -\sqrt{\mu L}R\Delta + \frac{L - \mu}{2}\Delta^2 \leq -\mu R \Delta. && (99)
\end{aligned}$$

Combining Equation (98) and Equation (99) completes our proof. □

Lemma 4.

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

$$\begin{aligned}
\mathbb{E}\|\hat{w}_m^{(k+1)} - w_m^*\|^2 &= \mathbb{E}\|\hat{w}_m^{(k)} - w_m^* - \eta_k g_m^{(k)}\|^2 && (100)\\
&= \mathbb{E}\|\hat{w}_m^{(k)} - w_m^* - \eta_k \hat{g}_m^{(k)}\|^2 + \eta_k^2 \mathbb{E}\|g_m^{(k)} - \hat{g}_m^{(k)}\|^2\\
&\qquad + 2\eta_k \mathbb{E}\langle w_m^{(k)} - w_m^* - \eta_k \hat{g}_m^{(k)}, \hat{g}_m^{(k)} - g_m^{(k)}\rangle && (101)\\
&= \mathbb{E}\|\hat{w}_m^{(k)} - w_m^* - \eta_k \hat{g}_m^{(k)}\|^2 + \eta_k^2 \mathbb{E}\|g_m^{(k)} - \hat{g}_m^{(k)}\|^2. && (102)
\end{aligned}$$

The first term can be decomposed into

$$\|\hat{w}_m^{(k)} - w_m^* - \eta_k \hat{g}_m^{(k)}\|^2 = \|\hat{w}_m^{(k)} - w_m^*\|^2 + \eta_k^2 \|\hat{g}_m^{(k)}\|^2 - 2\eta_k \langle \hat{w}_m^{(k)} - w_m^*, \hat{g}_m^{(k)}\rangle. \tag{103}$$

Note that

$$\|\hat{g}_m^{(k)}\|^2 \leq \sum_{e=1}^{E} \pi_{em} \|\nabla \mathcal{L}_e(\theta_{em}^{(k)})\|^2. \tag{104}$$

$$-\langle \hat{w}_m^{(k)} - w_m^*, \hat{g}_m^{(k)}\rangle = -\sum_{e=1}^{E} \pi_{em} \langle \hat{w}_m^{(k)} - \theta_{em}^{(k)}, \nabla \mathcal{L}_e(\theta_{em}^{(k)})\rangle - \sum_{e=1}^{E} \pi_{em} \langle \theta_{em}^{(k)} - w_m^*, \nabla \mathcal{L}_e(\theta_{em}^{(k)})\rangle. \tag{105}$$

We further decompose the two terms in Equation (105) by

$$-2\langle \hat{w}_m^{(k)} - \theta_{em}^{(k)}, \nabla \mathcal{L}_e(\theta_{em}^{(k)})\rangle \leq \frac{1}{\eta_k}\|\hat{w}_m^{(k)} - \theta_{em}^{(k)}\|^2 + \eta_k \|\nabla \mathcal{L}_e(\theta_{em}^{(k)})\|^2. \tag{106}$$

and

$$\begin{aligned}
\langle \theta_{em}^{(k)} - w_m^*, \nabla \mathcal{L}_e(\theta_{em}^{(k)})\rangle &\geq \langle \theta_{em}^{(k)} - w_m^*, \nabla \mathcal{L}_e(\theta_{em}^{(k)}) - \nabla \mathcal{L}_e(w_m^*)\rangle - \|\nabla \mathcal{L}_e(w_m^*)\|_2 \|\theta_{em}^{(k)} - w_m^*\|_2. && (107)\\
&\geq \frac{\mu L}{\mu + L}\|\theta_{em}^{(k)} - w_m^*\|^2 + \frac{1}{\mu + L}\|\nabla \mathcal{L}_e(\theta_{em}^{(k)} - \nabla \mathcal{L}_e(w_m^*))\|^2 - \|\nabla \mathcal{L}_e(w_m^*)\|_2 \|\theta_{em}^{(k)} - w_m^*\|_2. && (108)
\end{aligned}$$

Therefore,

$$
\begin{aligned}
\mathbb{E}\|\hat{w}_m^{(k+1)} - w_m^*\|^2 \le{} & \underbrace{\mathbb{E}\|\hat{w}_m^{(k)} - w_m^*\|^2 - 2\eta_k \frac{\mu L}{\mu + L} \sum_e \pi_{em} \mathbb{E}\|\theta_{em}^{(k)} - w_m^*\|^2}_{E_1} + \underbrace{\sum_e \pi_{em} \mathbb{E}\|\hat{w}_m^{(k)} - \theta_{em}^{(k)}\|^2}_{E_2} \\
& + \underbrace{\Big(2\eta_k^2 \sum_e \pi_{em} \mathbb{E}\|\nabla\mathcal{L}_e(\theta_{em}^{(k)})\|^2 - 2\eta_k \frac{1}{\mu + L} \sum_e \pi_{em} \mathbb{E}\|\nabla\mathcal{L}_e(\theta_{em}^{(k)}) - \nabla\mathcal{L}_e(w_m^*)\|^2\Big)}_{E_3} \\
& + \underbrace{2\eta_k \mathbb{E} \sum_e \pi_{em} \|\theta_{em}^{(k)} - w_m^*\|_2 \cdot \|\nabla\mathcal{L}_e(w_m^*)\|_2}_{E_4} + \underbrace{\eta_k^2 \mathbb{E}\|g_m^{(k)} - \hat{g}_m^{(k)}\|^2}_{E_5}.
\end{aligned}
\tag{109}
$$

$$
\begin{aligned}
E_1 &= \mathbb{E}\|\hat{w}_m^{(k)} - w_m^*\|^2 - 2\eta_k \frac{\mu L}{\mu + L} \mathbb{E}\Big(\sum_e \pi_{em} \|\hat{w}_m^{(k)} - w_m^*\|^2 + \sum_e \pi_{em} \|\hat{w}_m^{(k)} - \theta_{em}^{(k)}\|^2\Big) \\
&\le (1 - \frac{2\eta_k \mu L p \gamma_m}{\mu + L}) \mathbb{E}\|w_m^{(k)} - w_m^*\|^2 + E_2.
\end{aligned}
\tag{110}
$$

$$
\begin{aligned}
E_2 &= \mathbb{E} \sum_e \pi_{em} \|\hat{w}_m^{(k)} - \theta_{em}^{(k)}\|^2 \\
&= \mathbb{E} \sum_e \pi_{em} \|(w_m^{(0)} - \theta_{em}^{(k)}) + (\theta_{em}^{(k)} - w_m^{(k)})\|^2 \\
&\le \mathbb{E} \sum_e \pi_{em} \|(w_m^{(0)} - \theta_{em}^{(k)})\|^2 \\
&\le \sum_e \pi_{em} (K-1) \mathbb{E} \sum_{k'=0}^{k-1} {\eta_{k'}}^2 \|g_{em}(\theta_{em}^{(k')})\|^2 \\
&\le \frac{2\eta_k^2 E (K-1)^2 G^2 L^2}{\mu}.
\end{aligned}
\tag{111}
$$

$$
\begin{aligned}
E_3 &= 2\mathbb{E} \sum_e \pi_{em} \Big((\eta_k^2 - \frac{\eta_k}{\mu + L}) \|\nabla\mathcal{L}_e(\theta_{em}^{(k)})\|^2 + \frac{2\eta_k}{\mu + L} \langle \nabla\mathcal{L}_e(\theta_{em}^{(k)}), \nabla\mathcal{L}_e(w_m^*) \rangle - \eta_k \frac{\|\nabla\mathcal{L}_e(w_m^*)\|^2}{\mu + L} \Big) \\
&\le 4\eta_k \mathbb{E} \sum_e \pi_{em} \Big(- \frac{1}{2(\mu + L)} \|\nabla\mathcal{L}_e(\theta_{em}^{(k)})\|^2 + \frac{1}{\mu + L} \langle \nabla\mathcal{L}_e(\theta_{em}^{(k)}), \nabla\mathcal{L}_e(w_m^*) \rangle - \frac{\|\nabla\mathcal{L}_e(w_m^*)\|^2}{\mu + L} \Big) \\
&\le 6\eta_k \sum_e \pi_{em} \frac{\|\nabla\mathcal{L}_e(w_m^*)\|^2}{\mu + L} \\
&\le 6\eta_k \sum_{e \in S_m} \pi_{em} \frac{L^2 r^2}{\mu + L} + 6\eta_k \sum_{e \notin S_m} \pi_{em} \frac{2G}{\mu(\mu + L)} \\
&\le \eta_k O(r^2) + 6\eta_k \frac{G(1-p)E}{\mu(\mu + L)}.
\end{aligned}
\tag{112}
$$

$$
\begin{aligned}
E_4 &= 2\eta_k \mathbb{E} \sum_{e \in S_m} \pi_{em} \|\theta_{em}^{(k)} - w_m^*\|_2 \cdot \|\nabla\mathcal{L}_e(w_m^*)\|_2 + 2\eta_k \mathbb{E} \sum_{e \notin S_m} \pi_{em} \|\theta_{em}^{(k)} - w_m^*\|_2 \cdot \|\nabla\mathcal{L}_e(w_m^*)\|_2 \\
&\le 2\eta_k \gamma_m L r \sqrt{\frac{2G}{\mu}} + 2\eta_k (1-p) E L \cdot \frac{2G}{\mu}.
\end{aligned}
\tag{113}
$$

$$\begin{aligned}E_5 &= \eta_k^2 \mathbb{E}\|g_m^{(k)} - \hat{g}_m^{(k)}\|^2 \\ &\leq \eta_k^2 \mathbb{E}\Big\| \sum_e \pi_{em} \Big(\frac{1}{n_e} \sum_{i=1}^{n_e} \nabla \ell(h_{\theta_{em}}(x_e^{(i)}), y_e^{(i)}) - \mathcal{L}(\theta_{em}^{(k)}) \Big) \Big\|^2 \\ &\leq \eta_k^2 E \frac{\sigma^2}{n_e}. \end{aligned} \tag{114}$$

Combining Equation (110) to Equation (114) yields the conclusion of Lemma A.5. □

B Fairness Analysis

B.1 Proof of Theorem 4.6

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

$$\mathcal{E}_e(w_{m_1}) = \delta^2 \left\| \mu_e - \frac{\sum_{e'=1}^{E-1} \mu_{e'}}{E-1} \right\|^2 \leq \delta^2 r^2 \tag{115}$$

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

$$\mathcal{E}_E(w_{m_2}) = 0 \tag{116}$$

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

$$\mathcal{FAA}_{focus}(P) = \max_{i,j \in [E]} |\mathcal{E}_i(\Pi, W) - \mathcal{E}_j(\Pi, W)| \leq \delta^2 r^2. \tag{117}$$

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

$$\mathbb{E}_{(x,y)\sim\mathcal{P}_e}(\ell_{\hat{\theta}}(x)) = \mathbb{E}_{x\sim\mathcal{N}(0,\delta^2 I_d),\epsilon\sim\mathcal{N}(0,\sigma_e^2)}[(\mu_i^T x + \epsilon - \bar{\mu}^T x)^2] = \sigma_e^2 + \delta^2 \|\mu_e - \bar{\mu}\|^2 \tag{118}$$

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

$$\begin{align} \mathcal{E}_1(w_{avg}) &= \delta^2 \|\mu_1 - \bar{\mu}\|^2 \tag{119} \\ &= \delta^2 \|\mu_1 - \frac{\sum_{e=1}^{E-1} \mu_1}{E} - \frac{\mu_E}{E}\|^2 \tag{120} \\ &\leq \delta^2 \Big(r \cdot \frac{E-1}{E} + \frac{\|\mu_1 - \mu_E\|}{E} \Big)^2 \tag{121} \\ &\leq \delta^2 (r \cdot \frac{E-1}{E} + \frac{R+r}{E})^2 = \delta^2 (r + \frac{R}{E})^2 \tag{122} \end{align}$$

However for the outlier agent,

$$\begin{align} \mathcal{E}_E(w_{avg}) &= \delta^2 \|\mu_E - \bar{\mu}\|^2 \tag{123} \\ &= \delta^2 \left\| \frac{E-1}{E} \mu_E - \frac{\sum_{e=1}^{E-1} \mu_E}{E} \right\|^2 \tag{124} \\ &\geq \Big(\frac{E-1}{E} \Big)^2 \delta^2 R^2 \tag{125} \end{align}$$

Hence,

$$\mathcal{FAA}_{avg}(P) \geq \mathcal{E}_E(w_{avg}) - \mathcal{E}_1(w_{avg}) = \delta^2 \Big(\frac{R^2(E-2) - 2Rr}{E} + r^2 \Big) \tag{126}$$

□

Remark. When there are $E_k > 1$ outliers, we can similarly derive FAA for FedAvg algorithm:

$$\mathcal{E}_1(w_{avg}) \leq \delta^2 (r + \frac{E_k R}{E})^2 \tag{127}$$

$$\mathcal{E}_E(w_{avg}) \geq \delta^2 (\frac{E - E_k}{E} R - \frac{E_k}{E} r)^2 \tag{128}$$

so as long as $E_k < \frac{E}{2}$,

$$\mathcal{FAA}_{avg} \geq \mathcal{E}_E(w_{avg}) - \mathcal{E}_1(w_{avg}) = \Omega(\delta^2 R^2) \tag{129}$$

The FOCUS algorithm produces a result with

$$\mathcal{E}_1(w_{m_1}) \leq \delta^2 r^2 \tag{130}$$

$$\mathcal{E}_E(w_{m_2}) \leq \delta^2 r^2 \tag{131}$$

Hence we still have

$$\mathcal{FAA}_{focus} \leq \delta^2 r^2. \tag{132}$$

B.2 Proof of Theorem 4.7

We restate Theorem 4.7 for clarity here.

Theorem B.1. *The fairness FAA achieved by FOCUS with two clusters $M = 2$ is*

$$\mathcal{FAA}_{focus}(W, \Pi) \le \frac{2Gr}{E-1} \tag{133}$$

Let $B = \frac{2Gr}{E-1}$. The fairness achieved by FedAvg is

$$\mathcal{FAA}_{avg}(W) \ge \Big(\frac{E-1}{E} - \frac{L}{\mu E^2}\Big)R - \Big(1 + \frac{L(E-1)}{\mu E} - \frac{L^2}{\mu^2 E}\Big)B - \frac{2L}{\mu E}\sqrt{B(R - \frac{L}{\mu}B)} \tag{134}$$

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

$$\mathcal{L}_i(\theta) = \int p_i(x,y)\ell(f_\theta(x), y)\mathrm{d}x\mathrm{d}y. \tag{135}$$

Thus,

$$|\mathcal{L}_i(\theta_i^*) - \mathcal{L}_j(\theta_i^*)| = \int |p_i(x,y) - p_j(x,y)| \cdot \ell(f_{\theta_i^*}(x), y)\mathrm{d}x\mathrm{d}y \tag{136}$$

$$\le G \cdot \int |p_i(x,y) - p_j(x,y)|\mathrm{d}x\mathrm{d}y \le Gr. \tag{137}$$

Hence,

$$\mathcal{L}_i(\theta_j^*) \le \mathcal{L}_j(\theta_j^*) + Gr \le \mathcal{L}_j(\theta_i^*) + Gr \le \mathcal{L}_i(\theta_i^*) + 2Gr. \tag{138}$$

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

$$\mathcal{L}_i(\theta, \Pi) = \mathcal{L}_i\Big(\frac{\sum_{j=1}^{E-1}\theta_j^*}{E-1}\Big) \tag{139}$$

$$\le \frac{1}{T-1}\sum_{j=1}^{T-1}\mathcal{L}_i(\theta_j^*) \tag{140}$$

$$\le \mathcal{L}_i(\theta_i^*) + \frac{2Gr}{E-1}. \tag{141}$$

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

$$\mathcal{E}_i(\theta, \Pi) \le \frac{2Gr}{E-1}. \tag{142}$$

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

$$\mathcal{FAA}_{focus}(W, \Pi) \le \frac{2Gr}{E-1} \tag{143}$$

Now we prove the second part of Theorem 4.7 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

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

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

$$\mathcal{L}_E(\bar{\theta}') \ge \frac{1}{E-1}\Big(\sum_{i=1}^{E-1}\mathcal{L}_E(\theta_i^*) - \langle \nabla\mathcal{L}_E(\bar{\theta}'), \sum_{i=1}^{E-1}(\theta_i - \bar{\theta}')\rangle - \frac{L}{2}\sum_{i=1}^{E-1}\|\bar{\theta}' - \theta_i\|^2\Big) \tag{145}$$

$$= \frac{1}{E-1}\Big(\sum_{i=1}^{E-1}\mathcal{L}_E(\theta_i^*) - \frac{L}{2}\sum_{i=1}^{E-1}\|\bar{\theta}' - \theta_i\|^2\Big) \tag{146}$$

$$\ge \mathcal{L}_E(\theta_E^*) + R - \frac{LB}{\mu}. \tag{147}$$

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

$$B \geq \mathcal{L}_i(\bar{\theta}') - \mathcal{L}_i(\theta_i^*) \geq \frac{\mu}{2}\|\bar{\theta}' - \theta_i\|^2. \tag{148}$$

By L-smoothness, we have

$$\mathcal{L}_E(\bar{\theta}') \leq \mathcal{L}_E(\bar{\theta}) + \langle \nabla\mathcal{L}_E(\bar{\theta}), \bar{\theta}' - \bar{\theta}\rangle + \frac{L}{2}\|\bar{\theta}' - \bar{\theta}\|^2. \tag{149}$$

$$\mathcal{L}_E(\theta_E^*) \leq \mathcal{L}_E(\bar{\theta}) + \langle \nabla\mathcal{L}_E(\bar{\theta}), \theta_E^* - \bar{\theta}\rangle + \frac{L}{2}\|\theta_E^* - \bar{\theta}\|^2. \tag{150}$$

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

$$\mathcal{L}_E(\bar{\theta}) \geq \frac{(E-1)\mathcal{L}_E(\bar{\theta}') + \mathcal{L}_E(\theta_E^*) - \frac{L}{2}(E-1)\|\bar{\theta}' - \bar{\theta}\|^2 - \frac{L}{2}\|\theta_E^* - \bar{\theta}\|^2}{E} \tag{151}$$

$$= \frac{E-1}{E}\Big(R - \frac{LB}{\mu} - \frac{L\|\theta_E^* - \bar{\theta}'\|^2}{2E}\Big) + \mathcal{L}_E(\theta_E^*). \tag{152}$$

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

$$R - \frac{LB}{\mu} \geq \mathcal{L}_E(\bar{\theta}') - \mathcal{L}_E(\theta_E^*) \geq \frac{\mu}{2}\|\theta_E^* - \bar{\theta}'\|^2. \tag{153}$$

so

$$\mathcal{L}_E(\bar{\theta}) \geq (1 - \frac{L}{\mu E}) \cdot \frac{E-1}{E}(R - \frac{LB}{\mu}) + \mathcal{L}_E(\theta_E^*). \tag{154}$$

And

$$\mathcal{E}_E(\bar{\theta}) \geq (1 - \frac{L}{\mu E}) \cdot \frac{E-1}{E}(R - \frac{LB}{\mu}). \tag{155}$$

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

$$\mathcal{L}_i(\bar{\theta}) \leq \mathcal{L}_i(\bar{\theta}') + \langle \nabla\mathcal{L}_i(\bar{\theta}'), \bar{\theta} - \bar{\theta}'\rangle + \frac{L}{2}\|\bar{\theta} - \bar{\theta}'\|^2. \tag{156}$$

By L smoothness,

$$\|\nabla\mathcal{L}_i(\bar{\theta}')\|_2 \leq L\|\bar{\theta}' - \theta_i^*\| \leq L\sqrt{\frac{2B}{\mu}}. \tag{157}$$

So

$$\mathcal{L}_i(\bar{\theta}) \leq \mathcal{L}_i(\theta_i^*) + B + L\sqrt{\frac{2B}{\mu}}\sqrt{\frac{2(R - \frac{LB}{\mu})}{\mu}}\frac{1}{E} + \frac{L(R - \frac{LB}{\mu})}{\mu E^2} \tag{158}$$

$$\mathcal{E}_i(\bar{\theta}) \leq B + \frac{2L}{\mu E}\sqrt{B(R - \frac{LB}{\mu})} + \frac{L(R - \frac{LB}{\mu})}{\mu E^2} \tag{159}$$

In conclusion, the fairness can be estimated by

$$\mathcal{FAA}_{avg}(P) \geq \mathcal{E}_E(\bar{\theta}) - \mathcal{E}_1(\bar{\theta}) \tag{160}$$

$$\geq \Big(\frac{E-1}{E} - \frac{L}{\mu E^2}\Big)R - \Big(1 + \frac{L(E-1)}{\mu E} - \frac{L^2}{\mu^2 E}\Big)B - \frac{2L}{\mu E}\sqrt{B(R - \frac{L}{\mu}B)} \tag{161}$$

□

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 [45].

$$D_H(\mathcal{P}_e, \mathcal{P}_E) \geq \frac{LR}{4\mu} \tag{162}$$

Proof. Note that

$$\begin{aligned} D_H(\mathcal{P}_e, \mathcal{P}_E) &= \frac{1}{2}\min_\theta \Big(\mathcal{L}_e(\theta) + \mathcal{L}_E(\theta)\Big) + \frac{1}{2}\Big(\mathcal{L}_e(\theta_e^*) + \mathcal{L}_E(\theta_E^*)\Big) & (163)\\ &\leq \frac{1}{2}\Big(\mathcal{L}_e(\frac{\theta_e^* + \theta_E^*}{2}) + \mathcal{L}_E(\frac{\theta_e^* + \theta_E^*}{2})\Big) - \frac{1}{2}\Big(\mathcal{L}_e(\theta_e^*) + \mathcal{L}_E(\theta_E^*)\Big) & (164)\\ &\leq \frac{1}{2} \times (\frac{1}{2}L\|\frac{\theta_E^* - \theta_e^*}{2}\|_2^2 \times 2) & (165)\\ &= \frac{1}{8}L\|\theta_E^* - \theta_e^*\|_2^2 & (166) \end{aligned}$$

Therefore,

$$\begin{aligned} \mathcal{L}_E(\theta_e^*) - \mathcal{L}_E(\theta_E^*) &\geq \frac{\mu\|\theta_E^* - \theta_e^*\|_2^2}{2} & (167)\\ &\geq \frac{\mu}{2}\frac{8D_H(\mathcal{P}_e, \mathcal{P}_E)}{L} = R. & (168) \end{aligned}$$

□

C Fairness Discussion

This section discusses the difference between FAA and some existing fairness metrics.

We first formally recall the definition of existing fairness metrics. Suppose E clients join the federate learning process, and train models $\{\theta_1, \dots, \theta_E\}$. We denote the accuracy for client e as a_e.

$$\mathrm{Var}(a_1, \dots, a_E) \tag{Accuracy Parity}$$

$$\max_{e \in [E]} \mathcal{L}_e(\theta_e) \tag{Agnostic Loss}$$

$$\max_{e_1, e_2} \big((\mathcal{L}_{e_1}(\theta) - \min_w \mathcal{L}_{e_1}(w)) - (\mathcal{L}_{e_2} - \min_w \mathcal{L}_{e_2}(w))\big) \tag{FAA}$$

When the local data distributions of clients are identical, the Bayes optimal errors $\min_w \mathcal{L}_e(w)$ for all clients e are equal, which reduces FAA to

$$\max_{e_1, e_2 \in [E]} (\mathcal{L}_{e_1}(\theta_{e_1}) - \mathcal{L}_{e_2}(\theta_{e_2})). \tag{169}$$

This means when local data distributions are IID, optimizing agnostic loss is equivalent to optimizing our fairness metric FAA. In this case, agnostic loss is an upper bound of FAA.

However, both accuracy parity and agnostic loss malfunction when local distributions are disparate. An extreme but illustrative example is that a client e joins the federated learning with pure random noises as its data. The best prediction for this client is just a random guess, so $a_e = \frac{1}{C}$ (C is the number of classes in a classification task) and $\mathcal{L}_e(\theta_e)$ is high. On the other hand, FAA notifies the data contribution of client e by measuring its excess risk, i.e., $\mathcal{L}_e(\theta_e) - \min_w \mathcal{L}_e(w) = 0$. In conclusion, FAA is a generalization of agnostic loss to a more general scenario when local data distributions are heterogeneous.

D Additional Experimental Results

D.1 Experimental Setups

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 of 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	2	300
Yelp/IMDB	56000/25000	38000/25000	10	2	512	5e-5	2	3

D.2 Comparison with existing fair FL methods

We present the full results of existing fair federated learning algorithms on our data settings in terms of FAA. The results in Tables 3 and 4 show that FOCUS achieves the lowest FAA score compared to existing fair FL methods. We note that fair FL methods (i.e., q-FFL [22] and AFL [29]) have lower FAA scores than FedAvg, but their average test accuracy is worse. This is mainly because they mainly aim to improve bad agents (i.e., with high training loss), thus sacrificing the accuracy of agents with high-quality data.

Table 3: Comparison of FOCUS and the existing fair federated learning algorithms on the rotated MNIST dataset.

	FOCUS	FedAvg	q-FFL					AFL
			$q = 0.1$	$q = 1$	$q = 3$	$q = 5$	$q = 10$	$\lambda = 0.01$
Avg test accuracy	**0.953**	0.929	0.922	0.861	0.770	0.731	0.685	0.885
Avg test loss	**0.152**	0.246	0.269	0.489	0.777	0.900	1.084	0.429
FAA	**0.094**	0.363	0.388	0.612	0.547	0.419	0.253	0.220

Table 4: Comparison of FOCUS and the existing fair federated learning algorithms on the rotated CIFAR dataset.

	FOCUS	FedAvg	q-FFL					AFL
			$q = 0.1$	$q = 1$	$q = 3$	$q = 5$	$q = 10$	$\lambda = 0.01$
Avg test accuracy	**0.688**	0.654	0.648	0.592	0.426	0.181	0.121	0.661
Avg test loss	**1.133**	2.386	1.138	1.141	1.605	2.4746	2.526	1.666
FAA	0.360	1.115	0.620	0.473	0.384	**0.313**	0.379	0.595

D.3 Comparison with state-of-the-art FL methods

We compare FOCUS with other SOTA FL methods, including FedMA [35], Bayesian nonparametric FL [40] and FedProx [21]. Specifically, the matching algorithm in [40] is designed for only fully-connected layers, and the matching algorithm in [35] is designed for fully-connected and convolutional layers, while our experiments on CIFAR use ResNet-18 where the batch norm layers and residual modules are not considered in [35, 40]. Therefore, we evaluate [21, 35, 40] on MNIST with a fully-connected network, and [21] on CIFAR with a ResNet-18 model.

The results on MNIST and CIFAR in Tables 5 and 6 show that FOCUS achieves the highest average test accuracy and lowest FAA score than SOTA FL methods.

Table 5: Comparison of FOCUS and other SOTA federated learning algorithms on the rotated MNIST dataset.

	FOCUS	FedAvg	FedProx			FedMA	Bayesian Nonparametric
			$\mu = 1$	$\mu = 0.1$	$\mu = 0.01$		
Avg test accuracy	**0.953**	0.929	0.908	0.927	0.929	0.753	0.517
Avg test loss	**0.152**	0.246	0.315	0.252	0.246	0.856	2.293
FAA	**0.094**	0.363	0.526	0.378	0.365	1.810	0.123

Table 6: Comparison of FOCUS and other SOTA federated learning algorithms on the rotated CIFAR dataset.

	FOCUS	FedAvg	FedProx		
			$\mu = 1$	$\mu = 0.1$	$\mu = 0.01$
Avg test accuracy	**0.688**	0.654	0.647	0.643	0.653
Avg test loss	**1.133**	2.386	1.206	2.151	2.404
FAA	**0.360**	1.115	0.397	0.884	0.787

D.4 Scalability with more agents

To study the scalability of FOCUS, we evaluate the performance and fairness of FOCUS and existing methods under 100 clients on MNIST. Table 7 shows that FOCUS achieves the best fairness measured by FAA and Agnostic Loss, higher test accuracy, and lower test loss than Fedavg and existing fair FL methods.

Table 7: Comparison of different methods on MNIST 100 clients setting, in terms of average test accuracy (Avg Acc), average test loss (Avg Loss), fairness FAA and existing fairness metric Agnostic loss. FOCUS achieves the best fairness measured by FAA.

		FOCUS	FedAvg	q-FFL	AFL	Ditto	CGSV
				$q = 1$	$\lambda = 0.01$	$\lambda = 1$	$\beta = 1$
Rotated MNIST (100 clients)	Avg Acc	**0.9533**	0.9236	0.8371	0.8813	0.9351	0.8691
	Avg Loss	**0.157**	0.2571	0.5668	0.4355	0.2206	0.6294
	FAA	**0.5605**	1.0652	1.5055	0.8901	0.7459	1.2935
	Agnostic Loss	**0.5028**	0.8894	1.4227	0.7767	0.620	1.5133

D.5 Evaluation on scenarios with different numbers of outlier agents

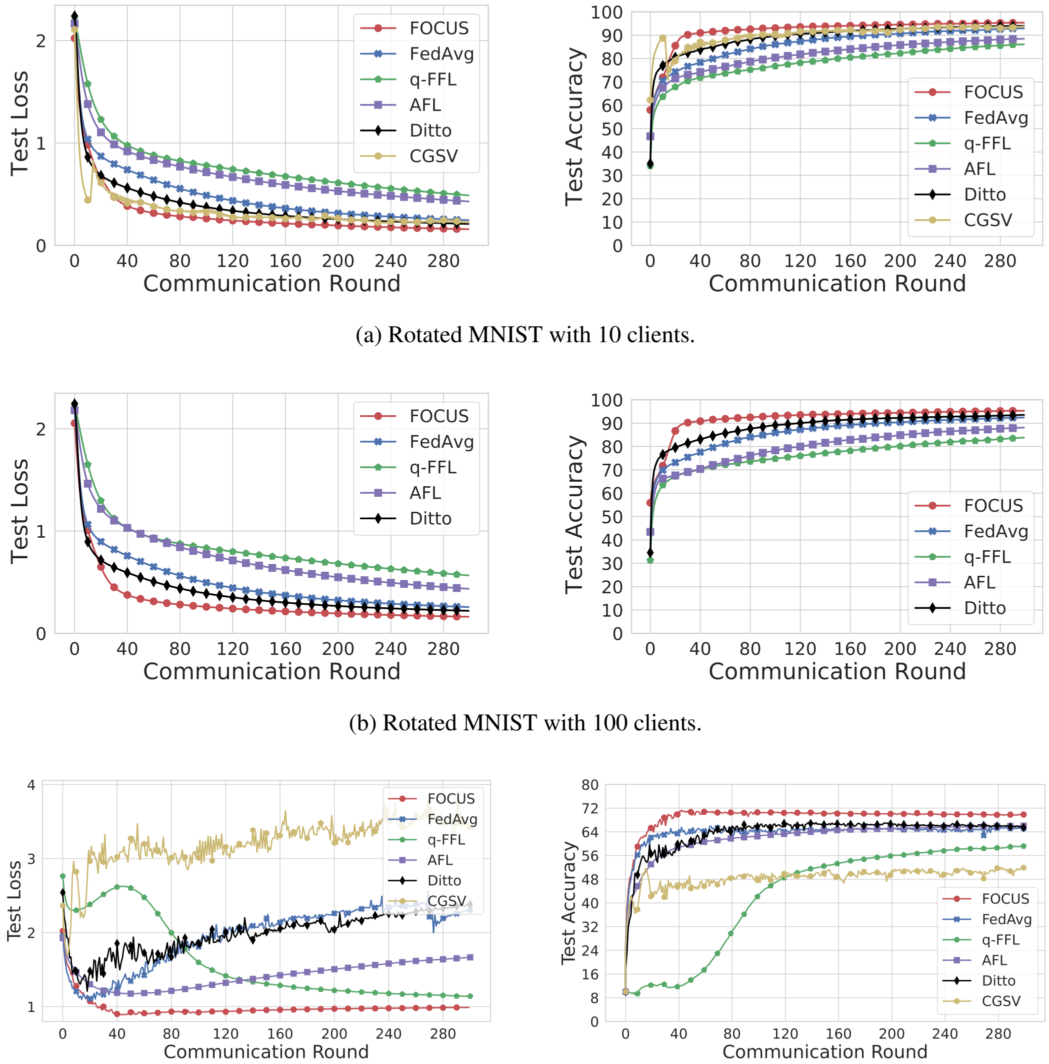
Additionally, we evaluate FOCUS with different numbers of outliers in Table 8. In the presence of 1, 3, and 5 outlier agents (e.g., agents with data rotated with 90 or 180 degrees), forming 2, 3, or 4 underlying true clusters, FOCUS consistently achieves a lower FAA score and higher accuracy.

Table 8: Comparison of FOCUS and FedAvg with different numbers of outlier agents (k) in terms of average test accuracy (Avg Acc) and fairness FAA.

		Rotated MNIST			Rotated CIFAR		
		$k = 1$	$k = 3$	$k = 5$	$k = 1$	$k = 3$	$k = 5$
Avg Acc	FOCUS	**0.957**	**0.953**	**0.948**	**0.683**	**0.688**	**0.677**
	FedAvg	0.945	0.929	0.910	**0.683**	0.654	0.651
FAA	FOCUS	**0.159**	**0.094**	**0.153**	**1.168**	**0.360**	**0.436**
	FedAvg	0.515	0.363	0.476	2.464	1.115	1.166

D.6 Convergence of FOCUS

We report the test accuracy and test loss of different methods over FL communication rounds on Rotated MNIST with 10/100 clients and Rotated CIFAR in Figure 2. The results show that FOCUS converges faster and achieves higher accuracy and lower loss than other methods on both settings.



(a) Rotated MNIST with 10 clients.

(b) Rotated MNIST with 100 clients.

(c) Rotated CIFAR with 10 clients.

Figure 2: The test accuracy and test loss of different methods over FL communication rounds on different datasets. FOCUS converges faster and achieves higher accuracy and lower loss than other methods.

D.7 Runtime analysis

Computation time analysis for proposed metric FAA and its scalability to more clients. In FAA, to calculate the maximal difference of excess risks for any pair of agents, it suffices to calculate the difference between the maximal per-client excess risks and the minimum per-client excess risk, and we don't need to calculate the difference for any pairs of agents. We compare the computation time (averaged over 100 trials) of FAA and existing fairness criteria (i.e., Accuracy Parity [22] and Agnostic Loss [29]) under 10 clients and 100 clients on MNIST. Table 9 shows that the computation of FAA is efficient even with a large number of agents. Moreover, calculating the difference between

Table 9: Computation time of different FL fairness metric on Rotated MNIST. The computation of FAA is efficient under 10 clients and 100 client settings.

	10 clients	100 clients
Accuracy Parity [22]	4.70 e-05 second	6.48 e-05 second
Agnostic loss [29]	9.41 e-07 second	3.92 e-06 second
FAA	6.09 e-06 second	4.27 e-05 second

Table 10: The number of communication rounds that different methods take to reach a target accuracy on Rotated MNIST. FOCUS requires a significantly smaller number of communication rounds than other methods.

	70%	80%	85%	90%
FOCUS	**9**	**16**	**20**	**29**
FedAvg	10	51	88	177
q-FFL	28	151	261	> 300
AFL	16	94	180	> 300

maximal excess risk and minimum excess risk (i.e., FAA) is even faster than calculating the standard deviation of the accuracy between agents (i.e., Accuracy Parity).

Communication rounds analysis. Here, we report the number of communication rounds that each method takes to achieve targeted accuracy on MNIST and CIFAR in Table 10. We note that FOCUS requires significantly a smaller number of communication rounds than FedAvg, q-FFL, and AFL on both datasets, which demonstrates the small costs required by FOCUS.

Training time and inference time analysis. In terms of runtime, we report the training time for one FL round (averaged over 20 trials) as well as inference time (averaged over 100 trials) in Table 12. Since the local updates and sever aggregation for different cluster models can be run in parallel, we find that FOCUS has a similar training time compared to FedAvg, q-FFL, and AFL which train one global FL model. For the inference time, FOCUS is slightly slower than existing methods by about 0.17 seconds due to the ensemble prediction of all cluster models at each client. However, we note that such cost is small and the forward passes of different cluster models for the ensemble prediction can also be made in parallel to further reduce the inference time.

D.8 Comparison to FedAvg with clustering

In this section, we construct a new method by combining the clustering and Fedavg together (i.e., FedAvg-HardCluster), which serves as a strong baseline. Specifically, FedAvg-HardCluster works as below:

Table 11: The number of communication rounds that different methods take to reach a target accuracy on Rotated CIFAR. FOCUS requires a significantly smaller number of communication rounds than other methods.

	55%	60%	65%	70%
FOCUS	**8**	**10**	**19**	**37**
FedAvg	**8**	14	34	> 300
q-FFL	182	> 300	> 300	> 300
AFL	24	53	190	> 300

Table 12: Training time per FL round and inference time for different methods on Rotate MNIST.

	Training time per FL round	Inference time
FOCUS	6.59 second	0.28 second
FedAvg	6.23 second	0.12 second
q-FFL	6.32 second	0.11 second
AFL	6.24 second	0.12 second

Table 13: Comparison between FOCUS and FedAvg-HardCluster on Rotate MNIST under two scenarios.

	Scenario 1 (underly clusters are clearly separatable)		Scenario 2 (underly clusters are not separatable)	
	FOCUS	FedAvg-HardCluster	FOCUS	FedAvg-HardCluster
Avg test acc	0.953	0.954	**0.814**	0.812
Avg test loss	0.152	0.152	**1.168**	1.244
FAA	**0.094**	0.099	**0.449**	0.459
Agnostic loss	0.224	0.224	**1.333**	1.397

- Step 1: before training, for each agent, it takes the arg max of the learned soft cluster assignment from FOCUS to get the hard cluster assignment (i.e., each agent only belongs to one cluster).
- Step 2: during training, each cluster then trains a FedAvg model based on corresponding agents.
- Step 3: during inference, each agent only uses the corresponding one cluster FedAvg model for inference.

To compare the performance between FOCUS and FedAvg-HardCluster, we consider two scenarios on MNIST:

- **Scenario 1**: underly clusters are clearly separatable, where each cluster contains samples from one distribution, which is the setting used in our paper.
- **Scenario 2**: underlying clusters are not separatable, where each cluster has 80%, 10%, and 10% samples from three different distributions, respectively. For example, the first underlying cluster contains 80% samples without rotation, 10% samples rotating 90 degrees, and 10% samples rotating 180 degrees.

We observe that the learned soft cluster assignments from FOCUS align with the underlying distribution, so the hard cluster assignment for Step 1 in FedAvg-HardCluster is equal to the underlying ground-truth clustering for both scenarios.

Table 13 presents the results of FOCUS and FedAvg-HardCluster on Rotated MNIST under two scenarios. **Under Scenario 1**, the accuracy of FOCUS and FedAvg-HardCluster is similar, and FOCUS achieves better fairness in terms of FAA. The results show that the hard clustering for FedAvg-HardCluster is as good as the soft clustering for FOCUS when the underlying clusters are clearly separable, which verifies that clustering is one of the key steps in FOCUS, and it aligns with our hypothesis for fairness under heterogeneous data. **Under Scenario 2**, FOCUS achieves higher accuracy and better FAA fairness than FedAvg-HardCluster. The results show that when underly clusters are not separatable, soft clustering is better than hard clustering since each agent can benefit from multiple cluster models with the soft π learned from the EM algorithm in FOCUS.

D.9 Effect of the number of the clusters M

The performance of FOCUS would not be harmed if the selected number of clusters is larger than the number of underlying clusters since the superfluous clusters would be useless (the corresponding soft cluster assignment π goes to zero). On the other hand, when the selected number of clusters is smaller than the number of underlying clusters, FOCUS would converge to a solution when some clusters contain agents from more than one underlying cluster.

Table 14: The effect of M on Rotate MNIST when the number of underlying clusters is 3.

	M=1	M=2	M=3	M=4
Avg test acc	0.929	0.952	**0.953**	**0.953**
Avg test loss	0.246	0.167	**0.152**	0.153
FAA	0.363	**0.079**	0.094	0.091
Agnostic loss	0.616	0.272	0.224	**0.223**

Table 15: The effect of M on Rotate CIFAR when the number of underlying clusters is 2.

	M=1	M=2	M=3	M=4
Avg test acc	0.654	0.688	**0.696**	0.693
Avg Loss	2.386	1.133	0.932	**0.921**
FAA	1.115	0.360	**0.323**	0.350
Agnostic loss	3.275	1.294	1.115	**1.098**

Rotate MNIST and Rotate CIFAR. Empirically, in Table 14, we have 3 true underlying clusters while we set $M = 1, 2, 3, 4$ in our experiments, and we see that when $M = 3$ and $M = 4$, FOCUS achieves similar accuracy and fairness, which verifies our hypothesis that the superfluous clusters would become useless. When $M = 2$, FOCUS even achieves the highest fairness, which might be because one cluster benefits from the shared knowledge of multiple underlying clusters. When $M = 1$, FOCUS reduces to FedAvg, which does not have the clustering mechanism, leading to the lowest accuracy and fairness under heterogeneous data.

FEMNIST. Here we conduct the experiment on FEMNIST, which represents a more realistic situation where local distributions are not well-separated. The FEMNIST contains 62 classes, and we consider 204 clients for FL, where each client contributes to a writer of handwritten digits or letters. The data distributions might exhibit ambiguous cluster structure since the writing style of different writers could potentially be clustered together.

We treat the number of clusters M as a hyperparameter and run the experiments with varying. The results in the below table show that, compared to FedAvg, FOCUS achieves higher accuracy and stronger fairness in terms of FAA and accuracy parity. Moreover, FOCUS is robust to the choice of the number of clusters as it maintains similar performance across $M = 2, 3, 4, 7, 10, 12, 15$. This indicates FOCUS is applicable in diverse FL settings, even when the cluster structure is ambiguous.

D.10 Histogram of loss on CIFAR

Figure 3 shows the surrogate excess risk of every agent trained with FedAvg and FOCUS on CIFAR dataset. For the outlier cluster that rotates 180 degrees (i.e., 2rd cluster), FedAvg has the highest test loss for the 9th agent, resulting in a high excess risk of 2.74. In addition, the agents in 1st cluster trained by FedAvg are influenced by the FedAvg global model and have high excess risk. On the other hand, FOCUS successfully identifies the outlier distribution in 2nd cluster, leading to a much lower excess risk among agents with a more uniform excess risk distribution. Notably, FOCUS reduces the surrogate excess risk for the 9th agent to 0.44.

Table 16: The effect of M on FEMNIST when the number of underlying clusters is unknown.

	M=1	M=2	M=3	M=4	M=7	M=10	M=12	M=15
Avg test acc	0.676	0.698	0.718	**0.727**	0.722	0.711	0.720	0.725
FAA	2.700	2.690	2.470	2.670	2.410	2.630	**2.370**	2.380
Accuracy parity	0.168	0.149	0.137	**0.135**	**0.135**	0.139	0.136	0.141

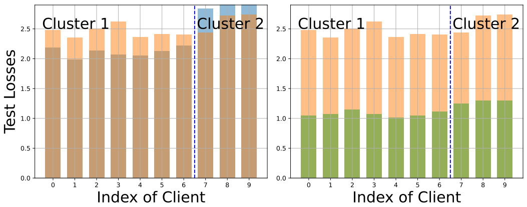


Figure 3: The excess risk of different agents trained with FedAvg (left) and FOCUS (right) on CIFAR dataset.

E Discussion

E.1 Limitation

A potential limitation of the FOCUS algorithm is that it requires setting the number of clusters M as a hyperparameter, as in many other FL clustering algorithms [13]. Our results on rotated MNIST, rotated CIFAR, and FEMINIST show that FOCUS is robust to the choice of M. Our future work includes evaluating FOCUS on more complex FL data distributions (e.g., real mobile devices data) and more data modalities (e.g., audio) and investigating the effect of clustering for fairness.

E.2 Broader Impact

This paper presents a novel definition of fairness via agent-level awareness for federated learning, which considers the heterogeneity of local data distributions among agents. 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.