

## A ADDITIONAL RELATED WORKS

**Personalized FL:** Personalized FL has received much attention. In addition to the clustering and local fine-tuning methods introduced in the main paper, we also review other types of personalization techniques used in FL framework.

**Model interpolation:** [Hanzely & Richtárik \(2020\)](#) also study a mixed model (local and global model) with a tuning parameter. In their model, as the mixing parameter decreases, it relaxes the local model to be similar to the global model, which can be more personalized. [Mansour et al. \(2020\)](#) propose an idea to combine the global and local model with weight  $\alpha$ , and [Deng et al. \(2020\)](#) adaptively find the optimal  $\alpha^*$  as a trade-off at each round for the best performance. [Zec et al. \(2020\)](#); [Peterson et al. \(2019\)](#) both consider using a gating model as a mixing parameter between local and global models. However, [Peterson et al. \(2019\)](#) consider a linear gating model and differentially private FL under domain adaptation, while [Zec et al. \(2020\)](#) split data into two parts used for local and global learning, and they further consider a dropout scenario and the same gating model structure as local and global models.

**Data interpolation:** As also suggested in [Mansour et al. \(2020\)](#), in addition to the model interpolation, it is possible to combine the local and global data and train a model on their combination. [Zhao et al. \(2018\)](#) create a subset of data that is globally shared across all clients. However, this method is facing the risk of information leaking.

**FL with Fairness.** In addition to works considering social (group) fairness in FL reviewed in the main paper, we review other types of FL fairness in detail below.

**Performance fairness:** This line of work measures fairness based on how well the learned model(s) can achieve uniform accuracy across all clients. [Li et al. \(2019a\)](#) propose the  $q$ -fair FL algorithm which minimizes the aggregate reweighted loss. The idea is that the clients with higher loss will be assigned a higher weight so as to encourage more uniform accuracy across clients. [Li et al. \(2021b\)](#) further extend this by considering robustness and poisoning attacks; here, performance fairness and robustness are achieved through a personalized FL method. [Zhang et al. \(2021\)](#) aim to achieve small disparity in accuracy across the groups of client-wise, attribute-wise, and potential clients with agnostic distribution, simultaneously. [Wang et al. \(2021\)](#) discuss the (performance) unfairness caused by conflicting gradients. They detect this conflict through the notion of cosine similarity, and iteratively eliminate it before aggregation by modifying the direction and magnitude of the gradients.

**Good-Intent fairness:** The good-intent fairness aims to minimize the maximum loss for the protected group. [Mohri et al. \(2019\)](#) propose a new framework of agnostic FL to mitigate the bias in the training procedure via minimax optimization. Similarly, [Cui et al. \(2021\)](#) consider a constrained multi-objective optimization problem to enforce the fairness constraint on all clients. They then maximize the worst client with fairness constraints through a gradient-based procedure. [Papadaki et al. \(2021\)](#) show that a model that is minimax fair w.r.t. clients is equivalent to a relaxed minimax fair model w.r.t. demographic group. They also show their proposed algorithm leads to the same minimax group fairness performance guarantee as the centralized approaches.

**Other types of fairness:** There are also other types of fairness considered in the FL literature. For instance, [Huang et al. \(2020\)](#) studied the unfairness caused by the heterogeneous nature of FL, which leads to the possibility of preference for certain clients in the training process. They propose an optimization algorithm combined with a double momentum gradient and weighting strategy to create a fairer and more accurate model. [Chu et al. \(2021\)](#) measure fairness as the absolute loss difference between protected groups and labels, a variant of equality opportunity fairness constraint. They propose an estimation method to accurately measure fairness without violating data privacy and incorporate fairness as a constraint to achieve a fairer model with high accuracy performance. Similarly, [Zhang et al. \(2022\)](#) study a new notion of fairness, proportional fairness, in FL, which is based on the relative change of each client’s performance. They connect with the Nash bargaining solution in the cooperative gaming theory and maximize the product of client utilities, where the total relative utility cannot be improved. Similarly, [Lyu et al. \(2020\)](#) study collaborative fairness, meaning that a client who has a higher contribution to learning should be rewarded with a better-performing local model. They introduce a collaborative fair FL framework that incorporates with reputation mechanism to enforce clients with different contributions converge to different models. Their approach could also be viewed as a variant of clustering that separates clients based on their contributions.

## B PERSONALIZATION CAN ALSO IMPROVE FAIRNESS: THEORETICAL SUPPORT

To support and validate our findings from the numerical experiments in Section 4, in this section, we analytically show that personalized Federated clustering algorithms (which cluster/group similar clients to improve their models' local accuracy) can also lead to better local fairness, when compared to a (non-personalized) shared global model.

We consider the following additional assumptions in our general model of Section 3. We assume the  $n$  clients can be potentially grouped into two clusters,  $C_\alpha$  and  $C_\beta$ , based on similarities in their data distributions  $f_g^{y,c}(x)$ , with a fraction  $p$  of clients in cluster  $C_\alpha$ .

We assume features are single dimensional  $x \in \mathbb{R}$ , and that clients can use their local data to learn a threshold-based, binary classifiers  $h_\theta(x) : \mathbb{R} \rightarrow \{0, 1\}$ <sup>1</sup> under which samples with features  $x \geq \theta$  are classified as label 1 (i.e.,  $\hat{y}(\theta) = 1$ ). Clients choose these thresholds to minimize classification errors. Formally, consider a client  $i$  from cluster  $c$ ; let  $r_g^c$  be the fraction of its samples that are from group  $g$ , and  $\alpha_g^{y,c}$  be the fraction of its samples that are from group  $g$  and have true label  $y$ . The client chooses its decision threshold  $\theta_i^*$  to (empirically) solve the following optimization problem:

$$\theta_i^* = \arg \min_{\theta} \sum_{g \in \{a,b\}} r_g^c \left( \alpha_g^{1,c} \int_{-\infty}^{\theta} f_g^{1,c}(x) dx + \alpha_g^{0,c} \int_{\theta}^{+\infty} f_g^{0,c}(x) dx \right). \quad (3)$$

For personalized learning, we consider a cluster-based FL algorithm where each cluster can learn its own optimal cluster-specific model  $\theta_i^*$ ,  $i \in \{\alpha, \beta\}$  (obtained after solving equation 3), and contrast that with the average optimal model  $\theta_G^*$  that would be obtained if all  $n$  clients collaboratively learn a shared global model. We then contrast the average local fairness  $\Delta_f^\alpha(\theta)$  obtained for clients in cluster  $C_\alpha$  under a personalized model  $\theta_\alpha^*$  vs. a shared model  $\theta_G^*$ , for two notions of fairness:  $f \in \{\text{EqOp}, \text{SP}\}$ .

We start with the EqOp (Equality of Opportunity) fairness constraint, which aims to equalize true positive rates (TPR) between the protected groups  $a$  and  $b$ . The following proposition shows that if  $\theta_\alpha^* < \theta_\beta^*$  (i.e., the data heterogeneity is such that cluster  $C_\alpha$  has a lower optimal threshold than  $C_\beta$ ), then clients in cluster  $C_\alpha$  can obtain better local fairness (in addition to better local accuracy) with their cluster-specific model compared to if they used a global model shared with clients in  $C_\beta$ .

**Proposition 1** (Improved EqOp through clustering). *Assume  $f_g^{y,c}(x)$ ,  $y \in \{0, 1\}$ ,  $g \in \{a, b\}$ ,  $c \in \{C_\alpha, C_\beta\}$ , are unimodal distributions, with modes  $m_g^{y,c}$  such that  $m_b^{y,c} \leq m_a^{y,c}$ ,  $\forall i, c$ , and  $\alpha_g^{1,c} \geq \alpha_g^{0,c}$ ,  $\forall g, c$ . If  $\theta_\alpha^* < \theta_\beta^*$ , there exist a cluster size  $\hat{p}$  such that for  $p \geq \hat{p}$ , we have  $\Delta_{\text{EqOp}}^\alpha(\theta_\alpha^*) < \Delta_{\text{EqOp}}^\alpha(\theta_G^*)$ ; that is, the global model is more unfair than the cluster-specific model for  $C_\alpha$ .*

The proof is presented in Appendix B.2.1. Intuitively, clients in  $C_\alpha$  are better off under their personalized model as, given the proposition's conditions, an increase in the decision threshold (which happens when moving from  $\theta_i^*$  to  $\theta_G^*$ ) will decrease the TPR of the disadvantaged group  $b$  (the one with a lower mode in its feature distribution) faster than that of the advantaged group  $a$ , increasing the fairness gap for clients in  $C_\alpha$ .

We next consider SP (Statistical Parity) fairness, which assesses to disparity in the selection (positive classification) rate between the two protected groups. This is impacted by both the group  $a$  vs.  $b$  feature distributions as well as the label rates, rendering it more stringent than EqOp fairness. Figure 9 illustrates this by plotting the fairness gap vs. the decision threshold  $\theta$  for SP vs. EqOp, showing that SP exhibits less structured changes as the decision threshold moves (e.g., due to the use of a global model). Therefore, to facilitate theoretical

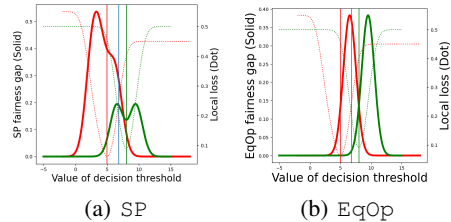


Figure 9: Fairness gap vs  $\theta$ .

<sup>1</sup>Our analysis assumes one-dimensional features and threshold classifiers. The former can be viewed as the one-dimensional representation of multi-dimensional features obtained from the last layer outputs of a neural network. For the latter, existing works (Corbett-Davies et al., 2017; Raab & Liu, 2021) show that threshold classifiers are optimal when multi-dimensional features can be properly mapped into a one-dimensional space.

analysis, we further assume that  $f_g^{y,c}(x)$  follows Gaussian distributions with equal variance  $\sigma^2$  but different means  $\mu_g^{y,c}$ , such that  $\mu_b^{0,c} \leq \mu_a^{0,c} \leq \mu_b^{1,c} \leq \mu_a^{1,c}$ ; the ordering is chosen so that label 0 samples have lower features than label 1 samples, and that for the same label, group  $a$  samples have higher features than group  $b$  samples (making group  $a$  advantaged). We again find that then clients in cluster  $C_\alpha$  can obtain better local SP-fairness (in addition to better local accuracy) with their cluster-specific model compared to if they joined in on a global model shared with clients in  $C_\beta$ .

**Proposition 2** (Improved SP through clustering). *Assume  $f_g^{y,c}(x), y \in \{0, 1\}, g \in \{a, b\}, c \in \{C_\alpha, C_\beta\}$ , are Gaussian distributions with means  $\mu_b^{0,c} \leq \mu_a^{0,c} \leq \mu_b^{1,c} \leq \mu_a^{1,c}$  and equal variance  $\sigma^2$ . Assume further that  $\mu_a^{1,c} - \mu_a^{0,c} = \mu_b^{1,c} - \mu_b^{0,c}$ ,  $\theta_\alpha^* < \theta_\beta^*$ , and that either  $\alpha_g^{1,c} \geq \alpha_g^{0,c}, \forall g$ , or  $r_b^c \geq r_a^c, \forall c$  and  $\alpha_a^{1,c} > \alpha_b^{0,c} > \alpha_b^{1,c} > \alpha_a^{0,c}$ . Then, if  $\alpha_a^{0,c} \exp(\frac{(\bar{\theta} - \mu_a^{0,c})^2}{-2\sigma^2})(\bar{\theta} - \mu_a^{0,c}) - \alpha_b^{1,c} \exp(\frac{(\bar{\theta} - \mu_b^{1,c})^2}{-2\sigma^2})(\bar{\theta} - \mu_b^{1,c}) \geq \alpha_b^{0,c} \exp(\frac{(\bar{\theta} - \mu_b^{0,c})^2}{-2\sigma^2})(\bar{\theta} - \mu_b^{0,c}) - \alpha_a^{1,c} \exp(\frac{(\bar{\theta} - \mu_a^{1,c})^2}{-2\sigma^2})(\bar{\theta} - \mu_a^{1,c})$  holds, where  $\bar{\theta} := \frac{\mu_a^{1,c} + \mu_b^{0,c}}{2}$ , there exist a  $\hat{p}$  such that for  $p \geq \hat{p}$ ,  $\Delta_{SP}^\alpha(\theta_\alpha^*) < \Delta_{SP}^\alpha(\theta_G^*)$ .*

A detailed proof is presented in Appendix B.2.2. Proposition 2 assumes an equal distance between mean estimates (reflecting a uniform, systematic underestimation of group  $b$  features); we relax this in Appendix D.1-D.4. Intuitively, the proposition states the following: when  $\alpha_g^{1,c} > \alpha_g^{0,c}$ , there are more label 1 data in both groups, and  $\theta_G^*$  will pull  $\theta_\alpha^*$  up to account for the label imbalance, resulting in a deterioration in both fairness and accuracy. Similarly, the other condition means that group  $a$ 's clients are majority label 1, while group  $b$ 's clients are majority label 0; then,  $\theta_\alpha^* < \bar{\theta}$  if  $r_b \geq r_a$ , resulting in a higher fairness gap for clients in  $C_\alpha$  under  $\theta_G^*$  for the same reason as the first condition.

## B.1 ADDITIONAL DISCUSSION ON THE ASSUMPTIONS USED IN THE ANALYTICAL RESULTS

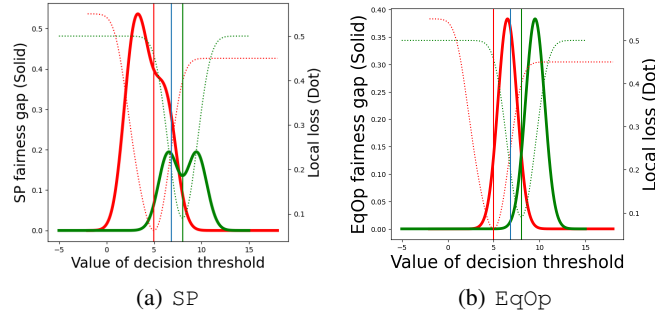


Figure 10: Fairness under different values of decision threshold

In Fig. 10, both cluster-wise SP and EqOp fairness and loss are assessed under the different values of decision threshold  $\theta$  on local data, depicted in red and green colors. The solid lines denote the corresponding fairness performance, while the dotted lines illustrate the corresponding loss under the corresponding decision threshold. The vertical red/green lines show the optimal decision threshold determined by solving Eq. 3, with the blue vertical line indicating the FedAvg solution with the parameter  $p = 0.5$ .

For a more insightful comparison, we assume that in the cluster  $C_\beta$  (green), the data distribution exhibits an equalized distance between distributions, equalized group rate, and equalized label rate (e.g.,  $(\mu_a^1, \mu_b^1, \mu_a^0, \mu_b^0) = (10, 9, 7, 6), \alpha_g^y = 0.5, r_g = 0.5$ ). Meanwhile, in the cluster  $C_\alpha$  (red), we relax all of these assumptions (e.g.,  $(\mu_a^1, \mu_b^1, \mu_a^0, \mu_b^0) = (7, 6, 4, 2), (\alpha_a^1, \alpha_b^1, \alpha_a^0, \alpha_b^0) = (0.6, 0.3, 0.4, 0.7), (r_a, r_b) = (0.65, 0.35)$ ). The comparison depicted in Fig. 10 reveals that the SP fairness exhibits less structured changes as the decision moves (e.g., due to the use of a global model). This is because EqOp fairness solely considers the true positive rates across two protected groups. The observation also underscores the necessity for more restrictive assumptions in our analytical support for SP fairness (Proposition 2).

## B.2 PROOFS

### B.2.1 PROOF OF PROPOSITION 1

For simplicity, we assume that clients within the same cluster are identical. Before we show the impact of the global model  $\theta_G^*$  on the fairness performance, we first prove that the global model will lie between two clusters' models  $\theta_\alpha^*$  and  $\theta_\beta^*$ .

**Lemma 1.** *Under the assumptions of our problem setup, the optimal solution  $\theta_G^*$  for the FedAvg algorithm will lie between  $\theta_\alpha^*$  and  $\theta_\beta^*$ .*

*Proof.* We prove this by contradiction. By definition, let  $\theta_G^* := \arg \min p * \sum_{j \in \mathcal{C}_\alpha} \mathcal{L}_j(\theta) + (1 - p) * \sum_{j \in \mathcal{C}_\beta} \mathcal{L}_j(\theta)$  and  $\theta_i^* := \arg \min \sum_{j \in \mathcal{C}_i} \mathcal{L}_j(\theta)$  are the optimal solutions for the FedAvg and the clustered FL algorithms, respectively, where  $\mathcal{L}_j$  is the objective function in Eq. 3. Without loss of generality, we assume  $\theta_\alpha^* < \theta_\beta^*$ . The following proof considers the scenario with  $\theta_G^* > \theta_\beta^*$ ; the other case can be shown similarly.

First, it is easy to verify that the objective function is convex in  $\theta$ . Then, if  $\theta_G^* > \theta_\beta^*$ , it should be that  $\sum_{j \in \mathcal{C}_\beta} \mathcal{L}_j(\theta_G^*) > \sum_{j \in \mathcal{C}_\beta} \mathcal{L}_j(\theta_\beta^*)$  because  $\theta_\beta^*$  can yield a smaller loss compared to the  $\theta_G^*$ . Similarly, we have  $\sum_{j \in \mathcal{C}_\alpha} \mathcal{L}_j(\theta_G^*) > \sum_{j \in \mathcal{C}_\alpha} \mathcal{L}_j(\theta_\beta^*) > \sum_{j \in \mathcal{C}_\alpha} \mathcal{L}_j(\theta_\alpha^*)$  due to convexity. Therefore,  $\theta_G^*$  is not the optimal solution, contradicting the assumption. Hence, the FedAvg solution would lie between  $\theta_\alpha^*$  and  $\theta_\beta^*$ .  $\square$

Now, we are ready to prove the Proposition 1 that global model is more unfair than the cluster-specific model (i.e.,  $\Delta_{\text{EqOp}}^\alpha(\theta_\alpha^*) < \Delta_{\text{EqOp}}^\alpha(\theta_G^*)$ ).

*Proof.* We start with the scenario where  $r_a = r_b$ , balanced label participation rates, and equalized distance between peaks. As the following analysis focuses on the cluster  $\mathcal{C}_\alpha$ , we drop the cluster notation from the derivation for notation simplicity. Let  $\Delta_{\text{EqOp}}(\theta)$  be the cluster-wise EqOp fairness gap at the given decision threshold  $\theta$ . Based on the definition, it could be written as

$$\Delta_{\text{EqOp}}(\theta) = \int_\theta^\infty f_a^1(x)dx - \int_\theta^\infty f_b^1(x)dx.$$

According to the Leibniz integral rule (Weisstein, 2003), we can find the derivative of  $\Delta_{\text{EqOp}}(\theta)$  w.r.t.  $\theta$  as following:

$$\Delta'_{\text{EqOp}}(\theta) = f_b^1(\theta) - f_a^1(\theta)$$

Let the intersection point of the feature-label distribution  $f_g^y$  and  $f_{g'}^{y'}$  be  $I_{g^y, g'^{y'}}$ . It is easy to verify that the optimal decision threshold  $\theta_\alpha^*$  obtained from 3 could be written in the closed form such that

$$\theta_\alpha^* = I_{a^1, b^0} = I_{b^1, a^0}$$

When  $\Delta'_{\text{EqOp}}(\theta) = 0$ ,  $\theta = \infty, -\infty$  or  $I_{a^1, b^1}$ . Furthermore, at extreme cases where  $\theta \rightarrow \infty$  or  $-\infty$ , we can find that the value of EqOp fairness gap  $\Delta_{\text{EqOp}}(\infty) = \Delta_{\text{EqOp}}(-\infty) = 0$ . Therefore, to investigate the impact of FedAvg solution  $\theta_G^*$  on the EqOp fairness gap, it is equivalent to check the sign of  $\Delta'_{\text{EqOp}}(\theta)$  at the optimal decision threshold  $\theta_\alpha^*$  obtained by solving 3.

To relax the equalized distance assumption, we could treat the location of modes of  $f_a^1, f_a^0, f_b^1$  as fixed, and vary the mode of  $f_b^0$ , and there are two cases we can discuss:

1.  $I_{a^1, b^1} - I_{a^0, b^1} < I_{a^0, b^1} - I_{a^0, b^0}$

Under this condition, we could consider a smaller value of the mode of  $f_b^0$ . As a result, the optimal decision threshold  $\theta_\alpha^*$  will shift to the left, resulting in a smaller value compared to the equalized distance case. In other words, it means  $\theta_\alpha^* \leq I_{a^1, b^1}$ , indicating  $\Delta'_{\text{EqOp}}(\theta_\alpha^*) \geq 0$ .

$$2. I_{a^1, b^1} - I_{a^0, b^1} > I_{a^0, b^1} - I_{a^0, b^0}$$

Under this condition, we could consider a larger value of the mode of  $f_b^0$ , but it is still less than that of  $f_a^0$  according to our assumption. At the extreme case when they are equal, the optimal decision threshold determined from [3] would be smaller than  $I_{a^0, a^1}$  because of the mode of  $f_b^1$  is less than that of  $f_a^1$ , which is also smaller than  $I_{a^1, b^1}$ . Therefore, we can still conclude  $\Delta'_{\text{EqOp}}(\theta_\alpha^*) \geq 0$ .

For the scenario of  $r_a \neq r_b$ , we can find that the change of  $r_g$  does not affect the value of  $\Delta_{\text{EqOp}}(\theta)$ , but the location of  $\theta_\alpha^*$ . According to our distribution assumption, when  $r_a \geq$  (resp.  $\leq$ )  $r_b$ , the optimal solution  $\theta_\alpha^*$  will be in favor of the group  $a$  (resp.  $b$ ) distributions, leading to a right (resp. left) shift compared to the optimal solution when  $r_a = r_b$ . However, when  $r_a \rightarrow 1$  (resp.  $0$ ),  $\theta_\alpha^* \rightarrow I_{a^0, a^1}$  (resp.  $I_{b^0, b^1}$ ), which is still less than  $I_{a^1, b^1}$ , indicating  $\Delta'_{\text{EqOp}}(\theta_\alpha^*) \geq 0$ .

With the assumption that the majority of samples are labeled as 1 (i.e.,  $\alpha_g^1 \geq \alpha_g^0$ ), the decision threshold  $\theta_\alpha^*$  will shift towards the left to account for label imbalance. In other words, the sign of  $\Delta'_{\text{EqOp}}(\theta_\alpha^*)$  remains positive. Since  $\theta_\alpha^* < \theta_\beta^*$ , there exist a cluster size weight  $p$  such that the FedAvg solution  $\theta_G^*$  will make the cluster  $\mathcal{C}_\alpha$  unfairer.  $\square$

## B.2.2 PROOF OF PROPOSITION 2

*Proof.* For simplicity, we assume that clients within the same cluster are identical. We start with the scenario where  $r_a = r_b$  and balanced label participation rate. As the following analysis focuses on the cluster  $\mathcal{C}_\alpha$ , we drop the cluster notation from the derivation for notation simplicity. Let  $\Delta_{\text{SP}}(\theta)$  be the cluster-wise SP fairness gap at the given decision threshold  $\theta$ . According to its definition, it could be written as

$$\Delta_{\text{SP}}(\theta) = \alpha_a^1 \int_\theta^\infty f_a^1(x) dx + \alpha_a^0 \int_\theta^\infty f_a^0(x) dx - \alpha_b^1 \int_\theta^\infty f_b^1(x) dx - \alpha_b^0 \int_\theta^\infty f_b^0(x) dx.$$

According to the Leibniz integral rule (Weisstein, 2003), we can find the derivative of  $\Delta_{\text{SP}}(\theta)$  w.r.t.  $\theta$  as following:

$$\Delta'_{\text{SP}}(\theta) = \alpha_b^1 f_b^1(\theta) + \alpha_b^0 f_b^0(\theta) - \alpha_a^1 f_a^1(\theta) - \alpha_a^0 f_a^0(\theta)$$

According to our distribution assumptions, we can write the above expression in the following closed form with  $\alpha = \alpha_y \forall y, g$

$$\Delta'_{\text{SP}}(\theta) = \frac{\alpha}{\sqrt{2\pi}\sigma} \left( \exp\left(-\frac{(\theta - \mu_b^1)^2}{2\sigma^2}\right) + \exp\left(-\frac{(\theta - \mu_b^0)^2}{2\sigma^2}\right) - \exp\left(-\frac{(\theta - \mu_a^1)^2}{2\sigma^2}\right) - \exp\left(-\frac{(\theta - \mu_a^0)^2}{2\sigma^2}\right) \right)$$

Furthermore, it is easy to verify that the optimal decision threshold  $\theta_\alpha^*$  obtained by solving [3] could be written in the closed form such that

$$\bar{\theta} = \theta_\alpha^* = \frac{\mu_a^1 + \mu_b^0}{2} = \frac{\mu_b^1 + \mu_a^0}{2}$$

At the optimal solution  $\theta_\alpha^*$ ,  $\Delta'_{\text{SP}}(\theta_\alpha^*) = 0$ . Similar to the proof of Proposition 1 to investigate the impact of FedAvg solution  $\theta_G^*$  on the SP fairness gap, it is equivalent to check how the  $\Delta'_{\text{SP}}(\theta_\alpha^*)$  change in the neighborhood of the optimal solution  $\theta_\alpha^*$ . Also, at extreme cases, we can easily find that the value of SP fairness gap  $\Delta_{\text{SP}}(\infty) = \Delta_{\text{SP}}(-\infty) = 0$ . Therefore, if  $\Delta'_{\text{SP}}(\theta_\alpha^*) \geq 0$ , then we can conclude that the FedAvg solution  $\theta_G^*$  would lead to a worse fairness performance compared to the optimal solution  $\theta_\alpha^*$ . Let  $\psi_1(\theta) = \exp\left(-\frac{(\theta - \mu_b^1)^2}{2\sigma^2}\right) - \exp\left(-\frac{(\theta - \mu_a^0)^2}{2\sigma^2}\right)$  and  $\psi_2(\theta) = \exp\left(-\frac{(\theta - \mu_a^1)^2}{2\sigma^2}\right) - \exp\left(-\frac{(\theta - \mu_b^0)^2}{2\sigma^2}\right)$ . At the solution  $\theta_\alpha^*$ , we can find that  $\psi_1(\theta_\alpha^*) = \psi_2(\theta_\alpha^*) = 0$ .

Hence, to investigate how the  $\Delta'_{\text{SP}}(\theta_\alpha^*)$  change, we can find the rate of change for both  $\psi_1(\theta)$  and  $\psi_2(\theta)$  in the neighborhood of  $\theta_\alpha^*$  such that

$$\begin{aligned} \psi'_1(\theta_\alpha^*) &= \exp\left(\frac{(\theta_\alpha^* - \mu_b^0)^2}{-2\sigma^2}\right) \frac{\theta_\alpha^* - \mu_b^0}{\sigma} - \exp\left(\frac{(\theta_\alpha^* - \mu_a^1)^2}{-2\sigma^2}\right) \frac{\theta_\alpha^* - \mu_a^1}{\sigma} = \frac{1}{\sigma} \exp\left(\frac{(\theta_\alpha^* - \mu_b^0)^2}{-2\sigma^2}\right) (\mu_b^1 - \mu_a^0) \\ \psi'_2(\theta_\alpha^*) &= \exp\left(\frac{(\theta_\alpha^* - \mu_b^0)^2}{-2\sigma^2}\right) \frac{\theta_\alpha^* - \mu_b^0}{\sigma} - \exp\left(\frac{(\theta_\alpha^* - \mu_a^1)^2}{-2\sigma^2}\right) \frac{\theta_\alpha^* - \mu_a^1}{\sigma} = \frac{1}{\sigma} \exp\left(\frac{(\theta_\alpha^* - \mu_b^0)^2}{-2\sigma^2}\right) (\mu_a^1 - \mu_b^0) \end{aligned}$$



By setting  $\psi'_1(\theta_\alpha^*) \geq \psi'_2(\theta_\alpha^*)$ , it means the increment of  $\psi_1$  is larger than the decrement of  $\psi_2$ . Therefore, with Lemma 1, there exists a cluster size weight  $p$  such that the FedAvg solution  $\theta_G^*$  will make the cluster  $\mathcal{C}_\alpha$  unfairer. The inequality is obtained by considering unequalized  $\alpha_g^y$ .

In addition, for the scenario of  $r_a \neq r_b$ , similar to the proof of Proposition 1, we can find that the change of  $r_g$  does not affect the expression of  $\Delta_{\text{SP}}(\theta)$ , but it will affect the location of  $\theta_\alpha^*$ . According to our distribution assumption, when  $r_a \geq$  (resp.  $\leq$ )  $r_b$ , the optimal solution  $\theta_\alpha^*$  will be in favor of the group  $a$  (resp.  $b$ ) distributions, leading to a right (resp. left) shift compared to the optimal solution when  $r_a = r_b$ . However, when  $r_a \rightarrow 1$  (resp.  $0$ ),  $\theta_\alpha^* \rightarrow \frac{\mu_a^0 + \mu_a^1}{2}$  (resp.  $\frac{\mu_b^0 + \mu_b^1}{2}$ ), which is limited within the range of  $(\frac{\mu_a^0 + \mu_b^0}{2}, \frac{\mu_a^1 + \mu_b^1}{2})$ . When  $\theta = \frac{\mu_a^1 + \mu_b^1}{2}$ , we can easily find that  $\Delta'_{\text{SP}}(\theta) \approx 0$  especially when  $\sigma$  is small. In other words, we can conclude that  $\Delta_{\text{SP}}(\theta) \geq \Delta_{\text{SP}}(\theta_\alpha^*)$  for any  $\theta_\alpha^* \in (\frac{\mu_b^0 + \mu_b^1}{2}, \frac{\mu_a^0 + \mu_a^1}{2})$ . Therefore, the claim still holds.

Furthermore, when the equalized label participation rate assumption is relaxed, the above proof strategy still holds by considering different  $\alpha_g^y$  into the expression. It is worth noting that when the label participation rates are balanced, the fairness  $\Delta_{\text{SP}}(\theta)$  has two equal-height peaks (e.g.,  $\Delta'_{\text{SP}}(\theta) = 0$ ) by symmetricity of the Gaussian distribution when  $\theta \approx \frac{\mu_a^1 + \mu_b^1}{2}$  and  $\frac{\mu_a^0 + \mu_b^0}{2}$ . However, when the majority of samples are labeled as 1, we observe a shift in the decision threshold  $\theta_\alpha^* \leq \bar{\theta}$  towards the left to account for label imbalance. In this case, since  $\theta_G^* > \theta_\alpha^*$ , the FedAvg solution pulls  $\theta_\alpha^*$  upwards, favoring label 1, which results in both accuracy and fairness deteriorating. Moreover, when  $r_b \geq r_a$  and the majority of samples are labeled 1 in one group where the other group has a better balance in the label (i.e.,  $\alpha_a^1 > \alpha_b^0 > \alpha_b^1 > \alpha_a^0$ ),  $\theta_\alpha^* \leq \bar{\theta}$  holds. Therefore,  $\Delta_{\text{SP}}(\theta)$  will increase initially and then decrease, and there still exist a cluster size weight  $p$  such that the FedAvg solution  $\theta_G^*$  will make the cluster  $\mathcal{C}_\alpha$  unfairer.  $\square$

## C EXPERIMENT DETAILS AND ADDITIONAL NUMERICAL EXPERIMENTS

### C.1 DATASET AND MODELS

In this section, we detail the data and model used in our experiments.

**Retiring Adult dataset.** We use the pre-processed dataset provided by the folktables Python package (Ding et al., 2021), which provides access to datasets derived from the US Census. In this package, there are three tasks: ACSEmployment, ACSIncome, and ACSHealth. For the ACSEmployment task, the goal is to predict whether the person is employed based on its multi-dimensional features; for the ACSIncome task, the goal is to predict whether the person earns more than \$50,000 annually; and for the ACSHealth task, the goal is to predict whether the person is covered by insurance.

**Model.** We train a fully connected two-layer neural network model for both tasks, where the hidden layer has 32 neurons for the ACSIncome task, and 64 neurons for the ACSEmployment and ACSHealth tasks. For all tasks, we use the RELU activation function and a batch size of 32. Furthermore, we utilize the SGD optimizer for training, with a learning rate of 0.001 for both FedAvg and MAML algorithms and 0.05 for the clustered FL algorithm. In FL, each client updates the global model for 10 epochs in the FedAvg and MAML algorithms and sends it back to the server, while the clustered FL algorithm that has a larger learning rate updates the global model for 1 epoch. We also follow the encoding procedure for categorical features provided by the folktables Python package. The input feature size is 54, 109 and 154 for the ACSIncome, ACSEmployment and ACSHealth tasks, respectively. In the experiments, we consider either sex (e.g., male and female) or race (e.g., White and Non-White) as the protected attribute.

#### ACSEmployment task with different protected attributes.

As shown in Figure 11 and 12, within the same ACSEmployment task, the data distributions for race (left) and sex (right) are significantly different. For the protected attribute of sex, the number of samples is nearly even across groups and labels. However, for the protected attribute of race, the White group has significantly more samples compared to Non-White groups for both labels 0, 1.

#### ACSIncome and ACSHealth tasks with protected attribute of sex.

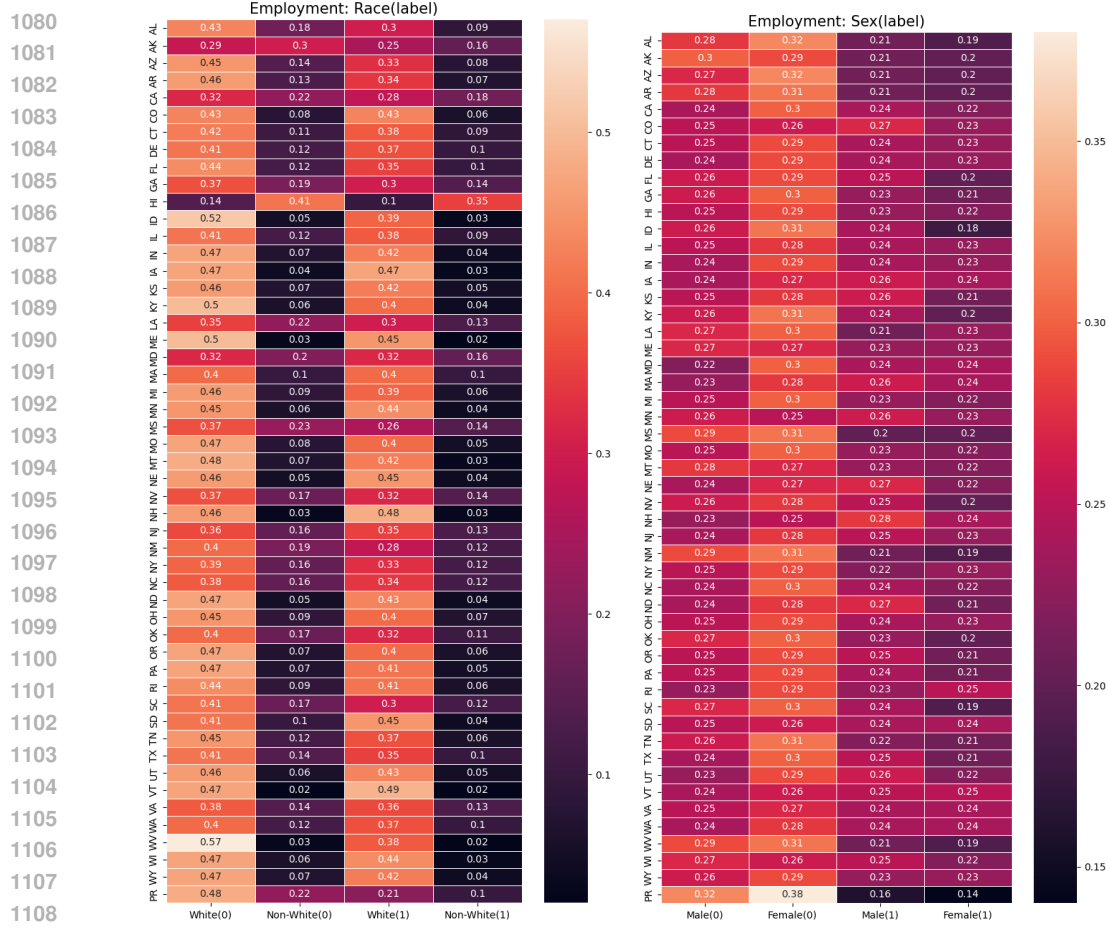


Figure 11: Fraction of samples over all states for ACSEmployment

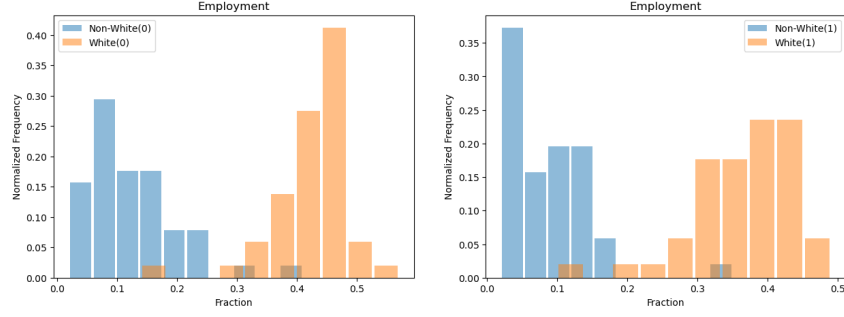
We can see from Figure 13, 14 and 15 that the fraction of samples in the ACSIncome task is similar across groups for label 0 data but differs significantly for label 1 data. Additionally, we can observe that the ACSHealth task has similar fractions of samples from each group, akin to the ACSEmployment task, in contrast to the ACSIncome task.

## C.2 ADDITIONAL EXPERIMENTS ON OTHER TYPES OF FAIRNESS NOTIONS

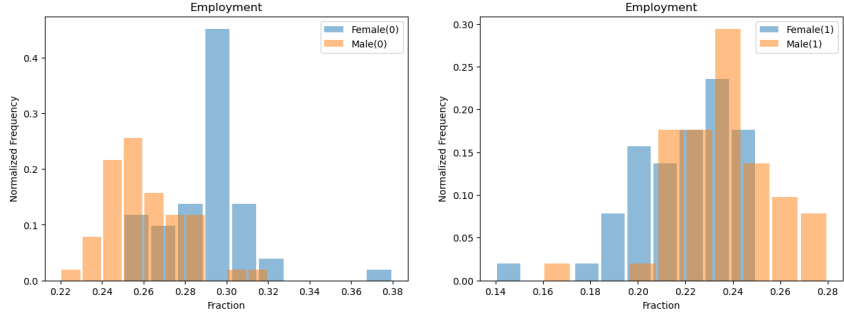
In addition to the SP fairness investigated in Section 4, we also study the impact of personalization techniques on other types of fairness notions such as EO and EqOp. From Fig. 16, we find that the introduction of personalization techniques can enhance other types of fairness due to the computational advantages of collaboration. However, compared to the improvement of SP and EqOp fairness, the local EO fairness improvement is less significant because the EO matches both the true and false positive rates across two protected groups, rendering it a more stringent criterion.

## C.3 ADDITIONAL EXPERIMENTS ON OTHER DATASET AND TASKS

In addition to the ACSEmployment (sex, race) and ACSIncome (sex) experimental results presented in Section 4, we conducted additional experiments to explore the impact of SP fairness using new datasets, as illustrated in Fig. 17. Examining the ACSHealth data with sex as the protected attribute, we can see from Fig. 15 that the fractions of samples from each group across all states are similar to that of the ACSEmployment dataset shown in Fig. 11, resulting in a similar performance. From Fig. 17, we can see that personalization techniques can improve local fairness as an unintended benefit, similar to the observations from Section 4.

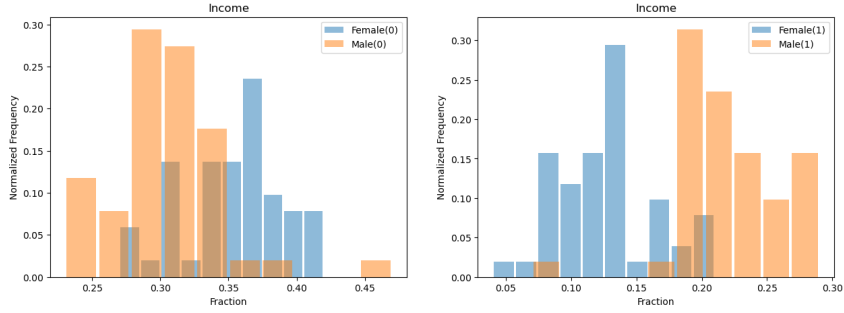


(a) Protected attribute: Race



(b) Protected attribute: Sex

Figure 12: Normalized frequency of fraction of samples for ACSEmployment



(a) Protected attribute: Sex

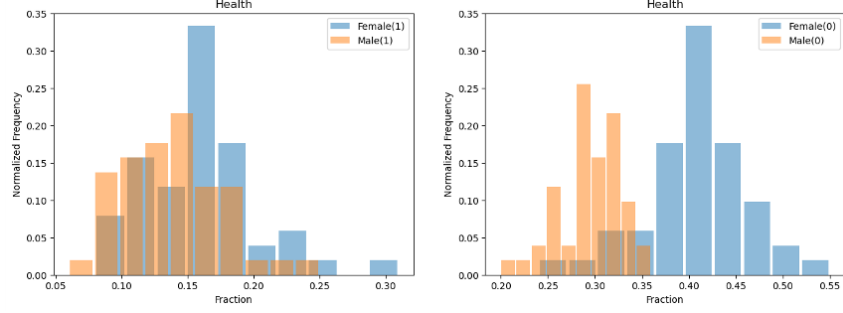
Figure 13: Normalized frequency of fraction of samples for ACSIncome

In the `Adult` dataset, where we randomly and evenly sample data into 5 clients. We could observe that the results are consistent with our findings in Section 4. That is, when groups are balanced (with sex as the protected attribute), the personalization could also improve the fairness as unintended benefit. However, when groups are unbalanced due to more White samples, the clustered FL algorithms have worse local fairness performance compared to `FedAvg`, but the `MAML-FL` algorithm could have a better performance.

#### C.4 ADDITIONAL EXPERIMENTS ON EQOP AND EO FAIRNESS

In Section 5, we compare SP fairness between two algorithms: `ICFA` and `Fair-FCA`. Here, we also compare the `EqOp` and `EO` fairness between them. The observations from Table 1 are also consistent with those from Fig 8, meaning that the `Fair-FCA` algorithm enables us to establish a better fairness-accuracy tradeoff (a drop in accuracy in return for improved fairness) compared to the `IFCA` algorithm.





(a) Protected attribute: Sex

Figure 14: Normalized frequency of fraction of samples for ACSHealth

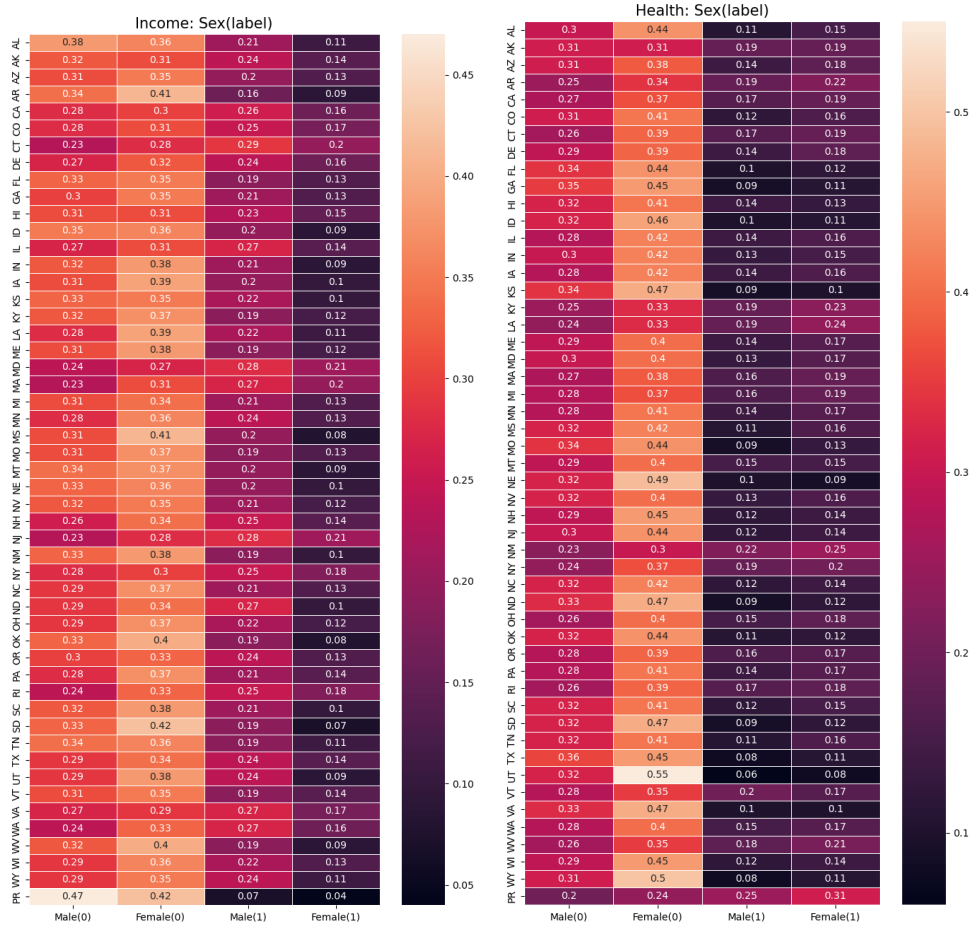


Figure 15: Fraction of samples over all states for ACSIncome and ACSHealth

### C.5 DETAILS OF SETUP ON SYNTHETIC EXPERIMENT

According to the data distribution information, we can see that clients 2,4,5,6,7,8 have similar data distributions compared to clients 1,3. Also, we can find that clients 1,3,4,6,7,8 share identical data distribution across the two groups. We generate 1200 samples from each distribution and apply a logistic regression classifier for binary classification tasks. We report our experiment results for an average of 5 runs. When  $\gamma = 1$ , Fair-FCA prioritizes accuracy; by design, this is attained by grouping the 6 clients having similar data distributions together ( $\{1,3\}$  and  $\{2,4,5,6,7,8\}$ ). Similarly, when  $\gamma = 0$ , Fair-FCA focuses only on SP fairness, this time clustering clients that have identical distributions on the two protected groups together ( $\{2,5\}$  and  $\{1,3,4,6,7,8\}$ ). Lastly, by setting

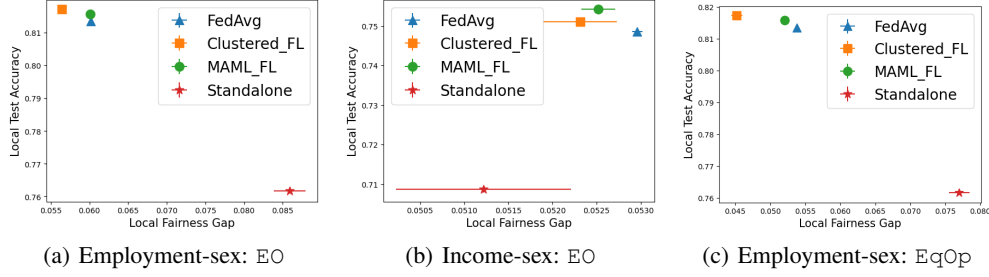


Figure 16: Personalization could also improve other fairness notions

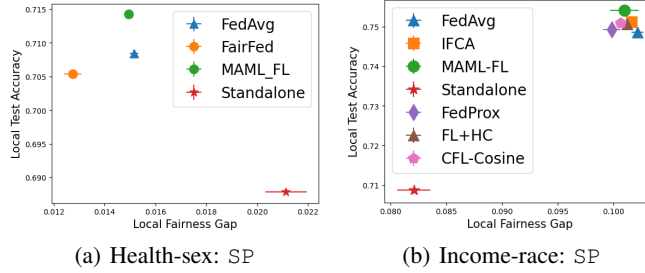


Figure 17: Additional experiments on other datasets with different protected attributes

$\gamma \in (0, 1)$ , we can effectively account for both accuracy and SP fairness when clustering: when  $\gamma = 0.3$ , the clusters are  $\{2, 4, 5\}$  and  $\{1, 3, 6, 7, 8\}$ ; when  $\gamma = 0.5$ , the clusters are  $\{2, 4, 5, 6\}$  and  $\{1, 3, 7, 8\}$ ; and when  $\gamma = 0.8$ , the clusters are  $\{2, 4, 5, 6, 7\}$  and  $\{1, 3, 8\}$ .

#### C.6 ADDITIONAL EXPERIMENTS USING ORIGINAL RETIRING ADULT DATASET WITHOUT FEATURE SCALING

We can see from Table 3 that compared to the IFCA algorithm, our Fair-FCA algorithm is experiencing a degradation in accuracy but an improved fairness, meaning an accuracy-fairness tradeoff. These observations are also consistent with our findings when using the Retiring Adult dataset with feature scaling in Section 5.

## D EXPERIMENTS ON SYNTHETIC DATA

To further validate our propositions, we conduct the following numerical experiments. In the experiments detailed in D.1 the setup is the most restrictive, with equalized distance, balanced group rates, and equalized label rates. In subsequent experiments, we relax one factor at a time. Finally, in the experiments described in D.4, all these assumptions are removed.

### D.1 EXPERIMENTS UNDER GAUSSIAN DISTRIBUTION WITH EQUALIZED DISTANCE, BALANCED GROUP RATE, AND EQUALIZED LABEL RATE

**Numerical illustration.** We now conduct numerical experiments to illustrate the findings in Prop. 1. We drop the cluster notation  $c$  whenever it is clear from the context. The results are presented in Tables 4 and 5. We proceed as follows: 10000 random samples in cluster  $C_\alpha$  are drawn from Gaussian distribution for each group  $g \in \{a, b\}$  with mean  $\mu_g^{y, C_\alpha}$  and standard deviation  $\sigma$ . The number of qualified ( $y = 1$ ) and unqualified ( $y = 0$ ) samples in each group is proportional to the label participation rate  $\alpha_g^{y, C_\alpha}$ . Since samples were generated in a consistent manner across different parameter settings, we assumed an optimal decision threshold  $\theta_\beta^* = 8$  for cluster  $C_\beta$ , obtained according to the distribution information:  $(f_1^1, f_1^0, f_0^1, f_0^0, \sigma) = (10, 7, 9, 6, 1)$  with equalized group rate  $r_g = 0.5, \forall g$  and label participation rate  $\alpha_g^{y, C_\beta} = 0.5, \forall g, y$ . In Table 4, we consider the scenario where  $\alpha_g^{y, C_\alpha} = 0.5 \forall g, y$ . In contrast, different values of  $\alpha_g^{y, C_\alpha}$  are applied in Table 5. Both results

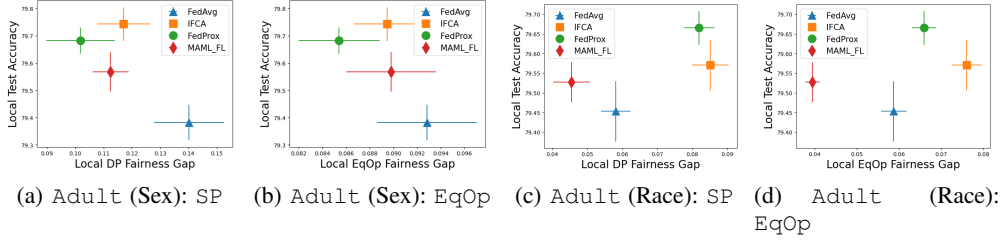


Figure 18: Additional experiments on Adult datasets with samples randomly and evenly distributed

Table 1: Fair-FCA with different datasets and protected attributes

Dataset	Algorithm	EqOp	Acc. (EqOp)	EO	Acc. (EO)
Employment-Race	IFCA	0.07764	0.8188	0.09319	0.8188
	Fair-FCA	0.07029	0.8151	0.08946	0.8151
Employment-Sex	IFCA	0.04521	0.8188	0.05808	0.8188
	Fair-FCA	0.04183	0.8157	0.05655	0.8151
Income-Sex	IFCA	0.05029	0.7511	0.05231	0.7511
	Fair-FCA	0.04932	0.7491	0.05161	0.7489

in Table 4 and 5 consider an equalized group rate such that  $r_a = r_b$  and an equalized distance between mean estimates.

From Table 4, we can find that it offers crucial insights into the conditions required for Proposition 2 (SP) to hold. For fixed mean estimates  $\mu_g^y$  (rows 1-2), we observe that smaller values of  $\sigma$  are preferable to satisfy the specified conditions. Similarly, for fixed  $\sigma$  (row 1, 3 and row 2, 4), larger differences between  $\mu_g^1$  and  $\mu_g^0$  are advantageous in fulfilling the conditions. This observation becomes intuitive at the extreme cases where samples are linearly separable with small  $\sigma$  or large distance between  $\mu_g^1$  and  $\mu_g^0$ . Therefore, the optimal decision threshold  $\theta_\alpha^*$  could achieve a perfect classification as well as perfect fairness. Hence, the FedAvg solution  $\theta_G^*$  deviated from the optimal solution will lead to worse performance in both accuracy and fairness. We could also observe that for the EqOp fairness, under an equalized label rate, the FedAvg solutions consistently make the cluster  $C_\alpha$  unfairer, which is consistent with our findings in Prop. 1.

Table 5 reveals insights regarding the influence of label distribution  $\alpha_g^y$  on SP and EqOp fairness performance. Specifically, when the majority of samples in both groups are labeled as 1 (rows 1-2), the optimal decision threshold ( $\theta_\alpha^*$ ) shifts leftward compared to the balanced scenario. However, with Lemma 1, the FedAvg solution  $\theta_G^*$  will be greater than  $\theta_\alpha^*$ . Therefore, we can find that  $\theta$  will have even larger fairness gap when it is shifted to the right. Another intriguing observation is that in cases where the majority of samples have different labels (row 3), the FedAvg solution ( $\theta_G^*$ ) yields worse fairness performance when  $p = 2/3$  or  $1/2$  but not when  $p = 1/3$  (0.1720  $\downarrow$ ) or  $1/4$  (0.1391  $\downarrow$ ). This indicates the weight  $p$  plays a significant role in shaping the overall cluster-wise average fairness performance, especially when assessing the overall cluster-wise average fairness performance.

Since we assume clients within the same cluster are identical, and the local fairness performance for an algorithm can be computed as a weighted sum of the local fairness performance from each cluster, the cluster-wise average local fairness gap under different models' optimal solution  $\theta$  could be calculated as  $\Delta_f(\theta) = p\Delta_f^\alpha + (1-p)\Delta_f^\beta$ ;  $f \in \{\text{SP}, \text{EqOp}, \text{EO}\}$ , where  $p$  is the fraction of clients belonging to cluster  $C_\alpha$ .

In Table 6 and 7 we delve into different notions of cluster-wise average fairness gap achieved with different decision thresholds (optimal clustered FL solutions  $\theta_C^*$  and FedAvg solutions  $\theta_G^*$ ). In the following experiment, we keep the parameters in cluster  $C_\beta$  as constants while varying those in cluster  $C_\alpha$  to assess its impact on the corresponding fairness. From the results in Table 6 and 7 we can find that when both conditions are not satisfied (rows 5-6), there is a cluster size weight  $p$  such

Table 2: Data distributions over 8 clients

Client ID	$f_1^1$	$f_1^0$	$f_0^1$	$f_0^0$
1	$N(8, 1)$	$N(6, 1)$	$N(8, 1)$	$N(6, 1)$
2	$N(12, 1)$	$N(8, 1)$	$N(11, 1)$	$N(7, 1)$
3	$N(7.5, 1)$	$N(5.5, 1)$	$N(7.5, 1)$	$N(5.5, 1)$
4	$N(12, 1)$	$N(9, 1)$	$N(12, 1)$	$N(9, 1)$
5	$N(12, 1)$	$N(8, 1)$	$N(11, 1)$	$N(7, 1)$
6	$N(11.5, 1)$	$N(8.5, 1)$	$N(11.5, 1)$	$N(8.5, 1)$
7	$N(11, 1)$	$N(8, 1)$	$N(11, 1)$	$N(8, 1)$
8	$N(10.5, 1)$	$N(7.5, 1)$	$N(10.5, 1)$	$N(7.5, 1)$

Table 3: Algorithm performance comparisons using original Retiring adult dataset

Dataset	Algorithm	SP	Acc. (SP)	EqOp	Acc. (EqOp)
Employment-Sex	IFCA	0.03667	0.8229	0.04698	0.8229
	Fair-FCA	0.03594 ↓	0.8223 ↓	0.04633 ↓	0.8224 ↓
Employment-Race	IFCA	0.07257	0.8229	0.07315	0.8229
	Fair-FCA	0.07219 ↓	0.8224 ↓	0.06527 ↓	0.8226 ↓
Income-Sex	IFCA	0.08355	0.7481	0.04773	0.7481
	Fair-FCA	0.08227 ↓	0.7469 ↓	0.04767 ↓	0.7469 ↓
Income-Race	IFCA	0.1012	0.7481	0.1100	0.7481
	Fair-FCA	0.1011 ↓	0.7468 ↓	0.1086 ↓	0.7466 ↓

that the FedAvg solutions would lead to better fairness performance for each cluster, consequently yielding a lower cluster-wise average fairness gap. However, when only one cluster satisfies the condition, meaning that there is a  $p$  such that the FedAvg solutions would only make one cluster unfairer (rows 1-2 in Table 6), we could see that a relatively small  $p$  would let the clustered FL solutions yield a better fairness performance because  $\theta_G^*$  will move to the cluster with a smaller value of  $p$  to account for the cluster size imbalance. Nevertheless, when  $p$  is large, the FedAvg solutions will again have superior fairness performance than the clustered FL solutions, similar to the results in rows 3-4 in Table 6 and 7. Essentially, for each cluster  $c$ , there exists a range  $(p_{low}^c, p_{high}^c)$  such that, within this range, FedAvg solutions result in worse fairness performance compared to clustered FL solutions. Consequently, for any  $p \in \cap_c (p_{low}^c, p_{high}^c)$ , clustered FL solutions yield a superior cluster-wise average fairness performance relative to FedAvg solutions.

## D.2 EXPERIMENTS UNDER GAUSSIAN DISTRIBUTION WITH EQUALIZED DISTANCE AND BALANCED LABEL RATE

Compared to the experiments focused on an all balanced setting in Table 4, the following experiments relax the group rates setting in the cluster  $\mathcal{C}_\alpha$ , while we keep other settings (i.e., balanced label rate and equalized distance) and data information for  $\mathcal{C}_\beta$  unchanged.

Table 4: Cluster  $\mathcal{C}_\alpha$  fairness performance with equalized distance, group rate and label rate

Distribution ( $\mu_a^1, \mu_a^0, \mu_b^1, \mu_b^0, \sigma$ )	Condition (SP)	$\Delta_{SP}^\alpha(\theta_\alpha^*)$	$p = \frac{2}{3}$	$p = \frac{1}{2}$	$\Delta_{EqOp}^\alpha(\theta_\alpha^*)$	$p = \frac{2}{3}$	$p = \frac{1}{2}$
(7, 4, 6, 3, 1)	Yes	0.1359	0.1814 $\uparrow$	0.1945 $\uparrow$	0.1359	0.3413 $\uparrow$	0.3829 $\uparrow$
(7, 4, 6, 3, 2)	No	0.1499	0.1417 $\downarrow$	0.1315 $\downarrow$	0.1499	0.1915 $\uparrow$	0.1974 $\uparrow$
(7, 5, 6, 4, 1)	No	0.2417	0.2297 $\downarrow$	0.2046 $\downarrow$	0.2417	0.3781 $\uparrow$	0.3721 $\uparrow$
(8, 3, 6, 1, 2)	Yes	0.1866	0.1968 $\uparrow$	0.2033 $\uparrow$	0.1866	0.3121 $\uparrow$	0.3590 $\uparrow$

Table 5: Cluster  $\mathcal{C}_\alpha$  fairness performance with equalized distance and group rate

Distribution ( $\mu_a^1, \mu_a^0, \mu_b^1, \mu_b^0, \sigma$ )	Label rate ( $\alpha_a^1, \alpha_a^0, \alpha_b^1, \alpha_b^0$ )	Condition (SP)	$\Delta_{SP}^\alpha(\theta_\alpha^*)$	$\Delta_{SP}^\alpha(\theta_G^*)$ $p = \frac{2}{3}$	$p = \frac{1}{2}$
(7, 4, 6, 3, 1)	(0.7, 0.3, 0.6, 0.4)	Yes	0.2062	0.2146 $\uparrow$	0.2463 $\uparrow$
	(0.6, 0.4, 0.7, 0.3)	Yes	0.0453	0.0514 $\uparrow$	0.0813 $\uparrow$
	(0.7, 0.3, 0.4, 0.6)	Yes	0.3797	0.3858 $\uparrow$	0.3926 $\uparrow$
	(0.6, 0.4, 0.3, 0.7)	No	0.3797	0.3748 $\downarrow$	0.2804 $\downarrow$
(7, 4, 6, 3, 2)		(EqOp)	$\Delta_{EqOp}^\alpha(\theta_\alpha^*)$	$\Delta_{EqOp}^\alpha(\theta_G^*)$	
	(0.7, 0.3, 0.6, 0.4)	Yes	0.0998	0.1807 $\uparrow$	0.1923 $\uparrow$
	(0.6, 0.4, 0.7, 0.3)	Yes	0.0975	0.1198 $\uparrow$	0.1796 $\uparrow$
	(0.1, 0.9, 0.5, 0.5)	No	0.1965	0.1650 $\downarrow$	0.1574 $\downarrow$
	(0.3, 0.7, 0.2, 0.8)	No	0.1974	0.1645 $\downarrow$	0.1574 $\downarrow$

From Table 8, we can see that the changes in the group rate do not affect the fairness performance comparison. There exists a cluster size weight  $p$  such that the FedAvg solutions would lead to worse SP and EqOp fairness performance compared to the clustered FL solutions. This observation is also consistent with our findings in the Proposition 1 and 2.

### D.3 EXPERIMENTS UNDER GAUSSIAN DISTRIBUTION WITH EQUALIZED DISTANCE

Similar to experiments in D.2, we further relax balanced label rate setting in the following experiments, while we keep other settings (i.e., equalized distance) and data information for  $\mathcal{C}_\beta$  unchanged.

From Table 9, we can observe that for the SP fairness, when the majority of samples are labeled 1 (rows 1-6), the changes in the group rate do not affect the fairness performance comparison in the cluster  $\mathcal{C}_\alpha$ . There exists a cluster size weight  $p$  such that the FedAvg solution would lead to a worse fairness performance compared to the clustered FL solutions. From Table 10, when the condition  $\alpha_g^1 \geq \alpha_g^0$  holds, there exists a combination of group rates (rows 1-6) such that the FedAvg solution would lead to a worse EqOp fairness performance. These observations from Table 9 and 10 are also consistent with our findings in the Proposition 1 and 2.

### D.4 ADDITIONAL EXPERIMENTS UNDER GAUSSIAN DISTRIBUTION

Similar to experiments in D.2 and D.3, we now release all settings we imposed before, while we data information for  $\mathcal{C}_\beta$  unchanged.

From Table 11, we can observe that when the majority of samples are labeled 1 (rows 1-3 and 7-9), there exists a cluster size weight  $p$  such that the FedAvg solution would lead to a worse SP fairness performance compared to the clustered FL solutions, which also experimentally extends our findings in the Proposition 2 to the case of an unequalized gap. However, when the majority of samples are labeled differently (rows 4-6 and 10-12), we could find that when  $\mu_a^1 - \mu_a^0 > \mu_b^1 - \mu_b^0$ , there exists a  $p$  such that the FedAvg solution would lead to a worse SP fairness performance, and a distinct outcome occurs when  $\mu_a^1 - \mu_a^0 < \mu_b^1 - \mu_b^0$ . One reason for the distinct behaviors is that the corresponding condition is not satisfied for the experiments in rows 4-6. Additionally, we find that as  $p$  enlarges in row 11, the fairness gap decreases, and it could have better fairness performance than using the clustered FL solution. This observation is also consistent with the previous finding that the fairness gap would increase initially and then decrease in the proof of Proposition 2. As we described earlier, it is clearly that for row 11,  $p = 1/2$  is not in the range of  $(p_{low}^{C_\alpha}, p_{high}^{C_\alpha})$ .

From Table 12, we could observe that when the condition  $\alpha_g^1 \geq \alpha_g^0$  holds (rows 1-3 and 7-9), the changes in the group rates, label rates, and distribution distance do not affect the EqOp fairness

Table 6: Cluster-wise average SP fairness performance with equalized distance

Distribution	Label rate	Condition	$p$	$\Delta_{\text{SP}}(\theta_C^*)$	$\Delta_{\text{SP}}(\theta_G^*)$
$\mathcal{C}_\alpha : (\mu_a^1, \mu_a^0, \mu_b^1, \mu_b^0, \sigma)$	$(\alpha_a^1, \alpha_a^0, \alpha_b^1, \alpha_b^0)$				
$\mathcal{C}_\beta : (\mu_a^1, \mu_a^0, \mu_b^1, \mu_b^0, \sigma)$	$(\alpha_a^1, \alpha_a^0, \alpha_b^1, \alpha_b^0)$	satisfied			
(7, 4, 6, 3, 2)	(0.5, 0.5, 0.5, 0.5)	No	4/5	0.147	0.145 ↓
(10, 7, 9, 6, 1)	(0.5, 0.5, 0.5, 0.5)	Yes	1/3	0.141	0.160 ↑
(7, 4, 6, 3, 2)	(0.8, 0.2, 0.7, 0.3)	Yes	3/4	0.139	0.107 ↓
(10, 7, 9, 6, 1)	(0.5, 0.5, 0.5, 0.5)	Yes	1/2	0.138	0.178 ↑
(7, 4, 6, 3, 2)	(0.5, 0.5, 0.5, 0.5)	No	1/3	0.303	0.283 ↓
(10, 7, 9, 6, 1)	(0.7, 0.3, 0.4, 0.6)	No	2/3	0.227	0.200 ↓

Table 7: Cluster-wise average EqOp fairness performance with equalized distance

Distribution	Label rate	Condition	$p$	$\Delta_{\text{EqOp}}(\theta_C^*)$	$\Delta_{\text{EqOp}}(\theta_G^*)$
$\mathcal{C}_\alpha : (\mu_a^1, \mu_a^0, \mu_b^1, \mu_b^0, \sigma)$	$(\alpha_a^1, \alpha_a^0, \alpha_b^1, \alpha_b^0)$				
$\mathcal{C}_\beta : (\mu_a^1, \mu_a^0, \mu_b^1, \mu_b^0, \sigma)$	$(\alpha_a^1, \alpha_a^0, \alpha_b^1, \alpha_b^0)$	satisfied			
(7, 4, 6, 3, 2)	(0.3, 0.7, 0.2, 0.8)	No	1/3	0.156	0.133 ↓
(10, 7, 9, 6, 1)	(0.5, 0.5, 0.5, 0.5)	No	2/3	0.177	0.139 ↓
(7, 4, 6, 3, 2)	(0.8, 0.2, 0.7, 0.3)	Yes	3/4	0.082	0.050 ↓
(10, 7, 9, 6, 1)	(0.5, 0.5, 0.5, 0.5)	No	1/2	0.100	0.109 ↑
(7, 4, 6, 3, 2)	(0.3, 0.7, 0.2, 0.8)	No	1/3	0.224	0.187 ↓
(10, 7, 9, 6, 1)	(0.3, 0.7, 0.2, 0.8)	No	2/3	0.211	0.149 ↓

performance in the cluster  $\mathcal{C}_\alpha$ . There exists a cluster size weight  $p$  such that the FedAvg solution would lead to a worse fairness performance. However, when the condition is not met (rows 4-6 and 10-12), the FedAvg solution would have a better EqOp fairness performance.



Table 8: Cluster  $\mathcal{C}_\alpha$  fairness performance under Gaussian distribution with equalized distance and label rate, but not group rate

Distribution ( $\mu_a^1, \mu_a^0, \mu_b^1, \mu_b^0, \sigma$ )	Group rate ( $r_a, r_b$ )	$\Delta_{\text{SP}}^\alpha(\theta_\alpha^*)$	$\Delta_{\text{SP}}^\alpha(\theta_G^*)$		$\Delta_{\text{EqOp}}^\alpha(\theta_\alpha^*)$	$\Delta_{\text{EqOp}}^\alpha(\theta_G^*)$	
			$p = \frac{2}{3}$	$p = \frac{1}{2}$		$p = \frac{2}{3}$	$p = \frac{1}{2}$
(7, 4, 6, 3, 1)	(0.5, 0.5)	0.1359	0.1814 ↑	0.1945 ↑	0.1359	0.3413 ↑	0.3829 ↑
	(0.7, 0.3)	0.1388	0.1558 ↑	0.1941 ↑	0.1780	0.2594 ↑	0.3828 ↑
	(0.9, 0.1)	0.1465	0.1702 ↑	0.1941 ↑	0.2217	0.3076 ↑	0.3828 ↑
	(0.3, 0.7)	0.1388	0.1359 ↓	0.1558 ↑	0.0996	0.1359 ↑	0.2594 ↑
	(0.4, 0.6)	0.1367	0.1372 ↑	0.1931 ↑	0.1161	0.1634 ↑	0.3759 ↑

Table 9: Cluster  $\mathcal{C}_\alpha$  SP fairness performance under Gaussian distribution with equalized distance, but not label rate and group rate

Distribution ( $\mu_a^1, \mu_a^0, \mu_b^1, \mu_b^0, \sigma$ )	Label rate ( $\alpha_a^1, \alpha_a^0, \alpha_b^1, \alpha_b^0$ )	Group rate ( $r_a, r_b$ )	$\Delta_{\text{SP}}^\alpha(\theta_\alpha^*)$	$\Delta_{\text{SP}}^\alpha(\theta_G^*)$	
				$p = \frac{2}{3}$	$p = \frac{1}{2}$
(7, 4, 6, 3, 1)	(0.7, 0.3, 0.6, 0.4)	(0.5, 0.5)	0.2062	0.2146 ↑	0.2463 ↑
		(0.3, 0.7)	0.2024	0.2056 ↑	0.2167 ↑
		(0.7, 0.3)	0.2136	0.2309 ↑	0.2793 ↓
	(0.6, 0.4, 0.7, 0.3)	(0.5, 0.5)	0.0453	0.0514 ↑	0.0813 ↑
		(0.3, 0.7)	0.0460	0.0446 ↓	0.0482 ↑
		(0.7, 0.3)	0.0535	0.0751 ↑	0.1467 ↑
	(0.7, 0.3, 0.4, 0.6)	(0.5, 0.5)	0.3797	0.3858 ↑	0.3926 ↑
		(0.3, 0.7)	0.3780	0.3819 ↑	0.3451 ↓
		(0.7, 0.3)	0.3821	0.3899 ↑	0.3936 ↑
	(0.4, 0.6, 0.7, 0.3)	(0.7, 0.3)	0.1005	0.0662 ↓	0.0766 ↓
		(0.9, 0.1)	0.0725	0.0078 ↓	0.0868 ↑
		(0.3, 0.7)	0.1013	0.1084 ↑	0.1090 ↓
		(0.1, 0.9)	0.0767	0.0860 ↑	0.0972 ↑

Table 10: Cluster  $\mathcal{C}_\alpha$  EqOp fairness performance under Gaussian distribution with equalized distance, but not label rate and group rate

Distribution ( $\mu_a^1, \mu_a^0, \mu_b^1, \mu_b^0, \sigma$ )	Label rate ( $\alpha_a^1, \alpha_a^0, \alpha_b^1, \alpha_b^0$ )	Group rate ( $r_a, r_b$ )	$\Delta_{\text{EqOp}}^\alpha(\theta_\alpha^*)$	$\Delta_{\text{EqOp}}^\alpha(\theta_G^*)$	
				$p = \frac{2}{3}$	$p = \frac{1}{2}$
(7, 4, 6, 3, 2)	(0.7, 0.3, 0.6, 0.4)	(0.5, 0.5)	0.0998	0.1222 ↑	0.1807 ↑
		(0.3, 0.7)	0.0952	0.1109 ↑	0.1784 ↑
		(0.7, 0.3)	0.1044	0.1386 ↑	0.1825 ↑
	(0.6, 0.4, 0.7, 0.3)	(0.5, 0.5)	0.0975	0.1198 ↑	0.1796 ↑
		(0.3, 0.7)	0.0798	0.0874 ↑	0.1799 ↑
		(0.7, 0.3)	0.1180	0.1957 ↑	0.1792 ↑
	(0.1, 0.9, 0.5, 0.5)	(0.5, 0.5)	0.1965	0.1650 ↓	0.1574 ↓
		(0.3, 0.7)	0.1751	0.1742 ↓	0.1620 ↓
		(0.7, 0.3)	0.1869	0.1569 ↓	0.1537 ↓
	(0.3, 0.7, 0.2, 0.8)	(0.5, 0.5)	0.1974	0.1645 ↓	0.1574 ↓
		(0.3, 0.7)	0.1974	0.1630 ↓	0.1569 ↓
		(0.7, 0.3)	0.1973	0.1660 ↓	0.1585 ↓

Table 11: Cluster  $\mathcal{C}_\alpha$  SP fairness performance under Gaussian distribution without equalized distance, label rate and group rate

Distribution ( $\mu_a^1, \mu_a^0, \mu_b^1, \mu_b^0, \sigma$ )	Label rate ( $\alpha_a^1, \alpha_a^0, \alpha_b^1, \alpha_b^0$ )	Group rate ( $r_a, r_b$ )	$\Delta_{\text{SP}}^\alpha(\theta_\alpha^*)$	$\Delta_{\text{SP}}^\alpha(\theta_G^*)$ $p = \frac{2}{3}$	$p = \frac{1}{2}$
(7, 4.5, 6, 3, 1)	(0.7, 0.3, 0.6, 0.4)	(0.5, 0.5)	0.2598	0.2649 ↑	0.2902 ↑
		(0.3, 0.7)	0.2589	0.2593 ↑	0.2655 ↑
		(0.7, 0.3)	0.2646	0.2781 ↑	0.3074 ↑
	(0.7, 0.3, 0.4, 0.6)	(0.5, 0.5)	0.4263	0.4220 ↓	0.3917 ↓
		(0.3, 0.7)	0.4288	0.4248 ↓	0.3222 ↓
		(0.7, 0.3)	0.4240	0.4198 ↓	0.3971 ↓
(7, 4, 6, 3.5, 1)	(0.7, 0.3, 0.6, 0.4)	(0.5, 0.5)	0.1871	0.2046 ↑	0.2483 ↑
		(0.3, 0.7)	0.1785	0.1910 ↑	0.2167 ↑
		(0.7, 0.3)	0.1984	0.2236 ↑	0.2784 ↑
	(0.7, 0.3, 0.4, 0.6)	(0.5, 0.5)	0.3576	0.3752 ↑	0.3882 ↑
		(0.3, 0.7)	0.3538	0.3697 ↑	0.3335 ↓
		(0.7, 0.3)	0.3620	0.3798 ↑	0.3903 ↑

Table 12: Cluster  $\mathcal{C}_\alpha$  EqOp fairness performance under Gaussian distribution without equalized distance, label rate and group rate

Distribution ( $\mu_a^1, \mu_a^0, \mu_b^1, \mu_b^0, \sigma$ )	Label rate ( $\alpha_a^1, \alpha_a^0, \alpha_b^1, \alpha_b^0$ )	Group rate ( $r_a, r_b$ )	$\Delta_{\text{EqOp}}^\alpha(\theta_\alpha^*)$	$\Delta_{\text{EqOp}}^\alpha(\theta_G^*)$ $p = \frac{2}{3}$	$p = \frac{1}{2}$
(7, 4.5, 6, 3, 2)	(0.7, 0.3, 0.6, 0.4)	(0.5, 0.5)	0.0993	0.1251 ↑	0.1796 ↑
		(0.3, 0.7)	0.0947	0.1115 ↑	0.1780 ↑
		(0.7, 0.3)	0.1045	0.1635 ↑	0.1814 ↑
	(0.3, 0.7, 0.2, 0.8)	(0.5, 0.5)	0.1948	0.1610 ↓	0.1558 ↓
		(0.3, 0.7)	0.1967	0.1610 ↓	0.1558 ↓
		(0.7, 0.3)	0.1924	0.1615 ↓	0.1558 ↓
(7, 4, 6, 3.5, 2)	(0.7, 0.3, 0.6, 0.4)	(0.5, 0.5)	0.1051	0.1409 ↑	0.1799 ↑
		(0.3, 0.7)	0.1016	0.1293 ↑	0.1776 ↑
		(0.7, 0.3)	0.1080	0.1564 ↑	0.1822 ↑
	(0.3, 0.7, 0.2, 0.8)	(0.5, 0.5)	0.1959	0.1625 ↓	0.1569 ↓
		(0.3, 0.7)	0.1950	0.1605 ↓	0.1553 ↓
		(0.7, 0.3)	0.1964	0.1650 ↓	0.1579 ↓