A ADDITIONAL RELATED WORKS

Personalized FL: Personalized FL has received much attention. The existing literature could be categorized based on how personalization is brought in.
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 α, and Deng et al. (2020) adaptively find the optimal α^* 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. The fairness in FL also gets lots of attention to ensure clients are treated fairly. Based on the type of fairness illustrated in the survey Shi et al. (2023); Rafi et al. (2023), it could also be categorized as follows:
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 equal 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 DATASET AND MODELS

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

Retiring adult. 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, it contains three tasks: ACSEmployment, ACSIncome, and ACSPublicCoverage. In this study, we focus on real-world experiments on the tasks of ACSEmployment and ACSIncome. For the ACSEmployemnt 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 could earn more than $50,000 annually.

Model. We train a fully connected two-layer neural network model for both tasks, where the hidden layer has 32 and 64 neurons for the Income and Employment tasks, respectively. For both tasks, we use the RELU activation function and a batch size of 32. Furthermore, we utilize the SGD optimizer for the 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 and 109 for the Income and Employment tasks respectively. Throughout the experiments, we consider both gender (e.g., male and female) and race (e.g., White and Non-White) as the protected attributes.

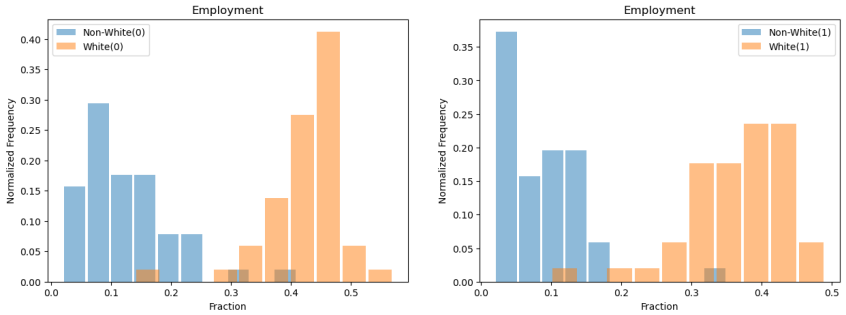
ACSEmployment dataset with different protected attributes.

Figure 5: Fraction of samples over all states for ACSEmployment

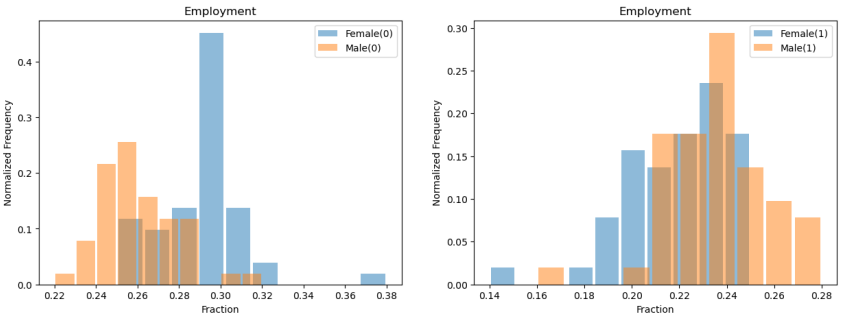
We could see from Figure 5 and 6 that by choosing different protected attributes: race (left) and gender (right), the group rates are significantly different even using the same ACSEmployment dataset. For the protected attribute gender, the amounts of samples are nearly even across groups and labels. However, for the protected attribute race, White groups has much more samples compared to Non-White groups in both labels $\{0, 1\}$.

ACSIncome dataset with gender as protected attributes.

We could see from Figure 7 and 8 that the fraction of samples is similar across groups for label 0 data, but differs significantly for label 1 data.

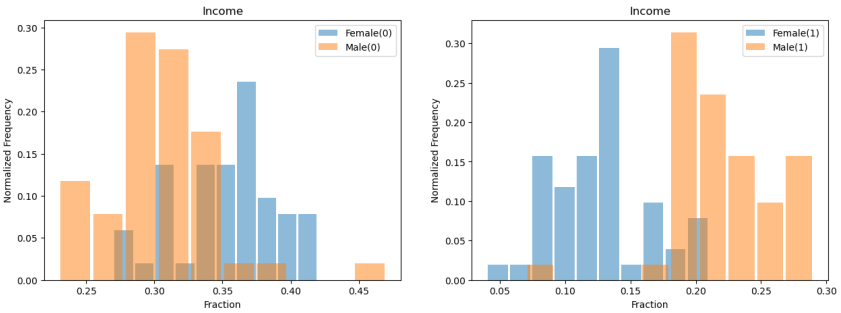


(a) Protected attribute: Race



(b) Protected attribute: Gender

Figure 6: Normalized frequency of fraction of samples for ACSEmployment



(a) Protected attribute: Gender

Figure 7: Normalized frequency of fraction of samples for ACSIncome

C Proofs

C.1 Proof of Proposition 1

Proof. Let $\Phi(\theta)$ be the cluster-wise statistical parity fairness gap at the given decision threshold θ. According to its definition, it could be written as

$$\Phi(\theta) = \alpha_a^1 \int_\theta^\infty f_a^1(x)\mathrm{d}x + \alpha_a^0 \int_\theta^\infty f_a^0(x)\mathrm{d}x - \alpha_b^1 \int_\theta^\infty f_b^1(x)\mathrm{d}x - \alpha_b^0 \int_\theta^\infty f_b^0(x)\mathrm{d}x.$$

According to the Leibniz integral rule (Weisstein, 2003), we can find the derivative of $\Phi(\theta)$ w.r.t. θ as following:

$$\Phi^{'}(\theta) = \alpha_b^1 f_b^1(\theta) + \alpha_b^0 f_b^0(\theta) - \alpha_a^1 f_a^1(\theta) - \alpha_a^0 f_a^0(\theta)$$

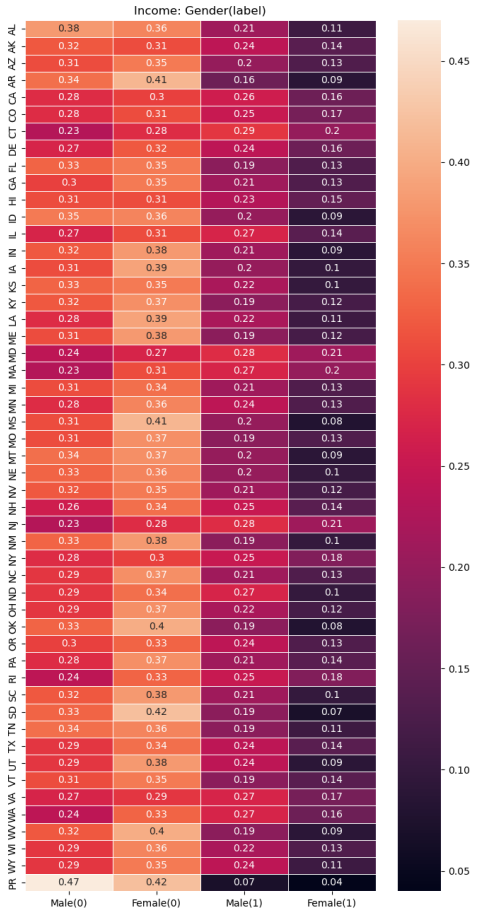


Figure 8: Fraction of samples over all states for ACSIncome

Based on our Gaussian distribution assumption with equal variance and equalized label participation rate α_g^y, we can write the above expression in the following closed form with $\alpha = \alpha_g^y \ \forall y, g$

$$\Phi^{'}(\theta) = \frac{\alpha}{\sqrt{2\pi}\sigma}\Big(\exp(-\frac{(\theta-\mu_b^1)^2}{2\sigma^2}) + \exp(-\frac{(\theta-\mu_b^0)^2}{2\sigma^2}) - \exp(-\frac{(\theta-\mu_a^1)^2}{2\sigma^2}) - \exp(-\frac{(\theta-\mu_a^0)^2}{2\sigma^2})\Big)$$

We start with the analysis where the $gr_a = gr_b$, it can be easily verified that at the scenario where the label participation rates are balanced, the optimal decision threshold $\theta^{*,\diamondsuit}$ obtained by solving 1 could be written in the closed form such that

$$\theta^{*,\diamondsuit} = \frac{\mu_a^1 + \mu_b^0}{2} = \frac{\mu_b^1 + \mu_a^0}{2}$$

At the optimal solution $\theta^{*,\diamondsuit}$, $\Phi^{'}(\theta^{*,\diamondsuit}) = 0$. To investigate the impact of FedAvg solution θ^{FA} on the statistical parity fairness gap, it is equivalent to check how the $\Phi^{'}(\theta^{*,\diamondsuit})$ change in the neighborhood of the optimal solution $\theta^{*,\diamondsuit}$. At extreme cases where $\theta \to \infty$ or $-\infty$, we can easily find that the value of statistical parity fairness gap $\Phi(\infty) = \Phi(-\infty) = 0$. Therefore, if $\Phi^{'}(\theta^{*,\diamondsuit}) \geq 0$, then we can conclude that the FedAvg solution θ^{FA} would lead to a larger fairness gap (worse fairness performance) compared to the optimal solution $\theta^{*,\diamondsuit}$. Let $\psi_1(\theta) = \exp(-\frac{(\theta-\mu_b^1)^2}{2\sigma^2}) - \exp(-\frac{(\theta-\mu_a^0)^2}{2\sigma^2})$ and $\psi_2(\theta) = \exp(-\frac{(\theta-\mu_a^1)^2}{2\sigma^2}) - \exp(-\frac{(\theta-\mu_b^0)^2}{2\sigma^2})$. At the solution $\theta^{*,\diamondsuit}$, we can find that $\psi_1(\theta^*) = \psi_2(\theta^*) = 0$.

Hence, to investigate how the $\Phi^{'}(\theta^{*,\diamondsuit})$ change, we can find the rate of change for both $\psi_1(\theta) = \psi_2(\theta)$ in the neighborhood of $\theta^{*,\diamondsuit}$ such that

$$\psi_1^{'}(\theta^{*,\diamondsuit}) = \exp(\frac{(\theta^{*,\diamondsuit}-\mu_a^0)^2}{-2\sigma^2})\frac{\theta^{*,\diamondsuit}-\mu_a^0}{\sigma} - \exp(\frac{(\theta^{*,\diamondsuit}-\mu_b^1)^2}{-2\sigma^2})\frac{\theta^{*,\diamondsuit}-\mu_b^1}{\sigma} = \frac{1}{\sigma}\exp(\frac{(\theta^{*,\diamondsuit}-\mu_b^1)^2}{-2\sigma^2})(\mu_b^1 - \mu_a^0)$$

$$\psi_2^{'}(\theta^{*,\diamondsuit}) = \exp(\frac{(\theta^{*,\diamondsuit}-\mu_b^0)^2}{-2\sigma^2})\frac{\theta^{*,\diamondsuit}-\mu_b^0}{\sigma} - \exp(\frac{(\theta^{*,\diamondsuit}-\mu_a^1)^2}{-2\sigma^2})\frac{\theta^{*,\diamondsuit}-\mu_a^1}{\sigma} = \frac{1}{\sigma}\exp(\frac{(\theta^{*,\diamondsuit}-\mu_b^0)^2}{-2\sigma^2})(\mu_a^1-\mu_b^0)$$

By setting $\psi_1^{'}(\theta^{*,\diamondsuit}) \geq \psi_2^{'}(\theta^{*,\diamondsuit})$, it means the increment of ψ_1 is larger than the decrement of ψ_2. Since $\theta^{*,\diamondsuit} < \theta^{*,\square}$, the FedAvg solution θ^{FA}, the weighted average of two clusters' optimal solutions, would be greater than $\theta^{*,\diamondsuit}$. Therefore, there exists a cluster size weight p such that $\theta^{FA} \geq \theta^{*,\diamondsuit}$. According to our analysis, we could also find that $\Phi(\theta^{FA}) > \Phi(\theta^{*,\diamondsuit})$. By plugging the closed form of $\theta^{*,\diamondsuit}$ into $\psi_1^{'}(\theta^{*,\diamondsuit}) \geq \psi_2^{'}(\theta^{*,\diamondsuit})$, it yields the required conditions.

Besides, for the scenario of $gr_a \neq gr_b$, we can find that the change of gr_g does not affect the statistical parity fairness gap $\Phi(\theta)$, but it will affect the location of $\theta^{*,\diamondsuit}$. According to our distribution assumption, when $gr_a \geq$ (resp. $\leq$)gr_b, the optimal solution $\theta^{*,\diamondsuit}$ will be in favor of the label 1 (resp. 0) distribution, leading to a right (resp. left) shift compared to the optimal solution when $gr_a = gr_b$. However, when $gr_a \to 1$ (resp. 0), $\theta^{*,\diamondsuit} \to \frac{\mu_a^0+\mu_a^1}{2}$ (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 $\Phi^{'}(\theta) \approx 0$ especially when σ is small. In other words, we can conclude that $\Phi(\theta) \geq \Phi(\theta^{*,\diamondsuit})$ for any $\theta^{*,\diamondsuit} \in (\frac{\mu_b^0+\mu_b^1}{2}, \frac{\mu_a^0+\mu_a^1}{2})$. Therefore, we are still able to draw the conclusion that there exists a p such that FedAvg solution θ^{FA} would lead to a worse fairness performance compared to the optimal decision threshold $\theta^{*,\diamondsuit}$ if $\psi_1^{'}(\theta^{*,\diamondsuit}) \geq \psi_2^{'}(\theta^{*,\diamondsuit})$.

□

C.2 Proof of Proposition 2

Proof. The proof technique is the same as the proof of Proposition 1. It is worth noting that when the label participation rates are balanced, the fairness $\Phi(\theta)$ has two equal-height peaks (e.g., $\Phi'(\theta) = 0$) by symmetricity of the Gaussian distribution when $\theta \approx \frac{\mu_a^1+\mu_b^1}{2}$ and $\frac{\mu_a^0+\mu_b^0}{2}$, especially when σ is small. However, when the majority of samples are labeled as 1, we observe a shift in the decision threshold $\theta^{*,\diamondsuit} \leq \bar{\theta}$ towards the left to account for label imbalance. In this case, since $\theta^{FA} > \theta^{*,\diamondsuit}$, the FedAvg solution pulls $\theta^{*,\diamondsuit}$ upwards, favoring label 1, which results in both accuracy and fairness deteriorating. Moreover, when the majority of samples are different between the two groups, the cluster experiences a worse accuracy-fairness trade-off. When $gr_b \geq gr_a$ and the majority of samples are labeled 1 in one group where the other group has a better balance in the label, $\theta^{*,\diamondsuit} \leq \bar{\theta}$ holds true. Under the specified conditions, $\Phi(\theta)$ will increase initially and then decrease. According to the assumption that $\theta^{*,\diamondsuit} < \theta^{*,\square}$, there exist a cluster size weight p such that the FedAvg solution $\theta^{FA} := p\theta^{*,\diamondsuit} + (1-p)\theta^{*,\square}$ will make the cluster $\diamondsuit$ unfairer. □

D Additional Numerical Experiments

D.1 Additional experiments under Gaussian distribution with equalized gap and balanced rate

The following experiments compared to the experiments in Table 1 consider different group rates in the cluster $\diamondsuit$, while we still assume the optimal decision threshold $\theta^{*,\square}$ to be 8, remains unchanged. In the following experiments, we only release the equalized group rate assumptions while the other assumptions (e.g., equalized gap and label rate) still hold.

In Table 4, we can see that under the assumptions of the equalized gap and label rate, the changes in the group rate do not affect the fairness performance in the cluster $\diamondsuit$. 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. This observation is also consistent with our findings in the Proposition 1.

D.2 Additional experiments under Gaussian distribution with equalized gap

The following experiments compared to the experiments in Table 2 consider different group rates in the cluster $\diamondsuit$, while we still assume the optimal decision threshold $\theta^{*,\square}$ to be 8, remains unchanged. In the following experiments, we release both the equalized group rate and label rate assumptions while the other assumptions (e.g., equalized gap) still hold.

Table 4: Cluster $\diamondsuit$ fairness performance under Gaussian distribution with equalized gap and label rate, but not group rate

Distribution $(\mu_a^1, \mu_a^0, \mu_b^1, \mu_b^0, \sigma)$	**Group rate** (gr_a, gr_b)	$ASPD_{\diamondsuit}(\theta^{*,\diamondsuit})$	$ASPD_{\diamondsuit}(\theta^{FA})$ $p=\frac{2}{3}$	$p=\frac{1}{2}$
(7, 4, 6, 3, 1)	(0.5, 0.5)	0.1359	0.1814 $\uparrow$	0.1945 $\uparrow$
	(0.7, 0.3)	0.1388	0.1881 $\uparrow$	0.1927 $\uparrow$
	(0.9, 0.1)	0.1465	0.1925 $\uparrow$	0.1895 $\uparrow$
	(0.2, 0.8)	0.1421	0.1691 $\uparrow$	0.1939 $\uparrow$
	(0.4, 0.6)	0.1367	0.1773 $\uparrow$	0.1947 $\uparrow$

Table 5: Cluster $\diamondsuit$ fairness performance under Gaussian distribution with equalized gap, 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** (gr_a, gr_b)	$ASPD_{\diamondsuit}(\theta^{*,\diamondsuit})$	$ASPD_{\diamondsuit}(\theta^{FA})$ $p=\frac{2}{3}$	$p=\frac{1}{2}$
(7, 4, 6, 3, 1)	(0.7, 0.3, 0.6, 0.4)	(0.5, 0.5)	0.2062	0.2832 $\uparrow$	0.3041 $\uparrow$
		(0.3, 0.7)	0.2024	0.2740 $\uparrow$	0.3043 $\uparrow$
		(0.7, 0.3)	0.2136	0.2921 $\uparrow$	0.3021 $\downarrow$
	(0.6, 0.4, 0.7, 0.3)	(0.5, 0.5)	0.0453	0.1433 $\uparrow$	0.1961 $\uparrow$
		(0.3, 0.7)	0.0460	0.1236 $\uparrow$	0.1886 $\uparrow$
		(0.7, 0.3)	0.0535	0.1627 $\uparrow$	0.2012 $\uparrow$
	(0.7, 0.3, 0.4, 0.6)	(0.5, 0.5)	0.3797	0.3962 $\uparrow$	0.3676 $\downarrow$
		(0.3, 0.7)	0.3780	0.3969 $\uparrow$	0.3734 $\downarrow$
		(0.7, 0.3)	0.3821	0.3945 $\uparrow$	0.3607 $\downarrow$
				$p=\frac{999}{1000}$	$p=\frac{2}{3}$
(7, 4, 6, 3, 1)	(0.4, 0.6, 0.7, 0.3)	(0.7, 0.3)	0.1005	0.1003 $\downarrow$	0.0205 $\downarrow$
		(0.9, 0.1)	0.0725	0.0722 $\downarrow$	0.0462 $\downarrow$
		(0.3, 0.7)	0.1013	0.1015 $\uparrow$	0.0367 $\downarrow$
		(0.1, 0.9)	0.0767	0.0771 $\uparrow$	0.0632 $\downarrow$

From Table 5, we can observe that under the assumptions of the equalized gap, when the majority of samples are labeled 1 (rows 1-6), the changes in the group rate do not affect the fairness performance in the cluster $\diamondsuit$. 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. When the majority of samples are labeled differently (rows 7-13), we can find that there always exists a p such that the FedAvg solution would lead to a worse fairness performance when $gr_a \leq gr_b$. However, when the $gr_a \geq gr_b$, the FedAvg could lead to a worse fairness performance (row 9) or a better fairness performance (rows 11-12) for certain cluster size weight p. This observation is also consistent with our findings in the Proposition 2.

D.3 Additional experiments under Gaussian distribution

The following experiments compared to the experiments in Table 2 consider different group rates in the cluster $\diamondsuit$, while we still assume the optimal decision threshold $\theta^{*,\square}$ to be 8, remains unchanged. In the following experiments, we release all the assumptions of equalized group rate, label rate, and gap.

From Table 6, 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 fairness performance compared to the clustered FL solutions under the unequalized gap scenario (e.g., $\mu_a^1 - \mu_a^0 \neq \mu_b^1 - \mu_b^0$). This observation also experimentally extends our findings in the Proposition 2.

Table 6: Cluster $\diamondsuit$ fairness performance under Gaussian distribution with equalized gap, 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** (gr_a, gr_b)	$ASPD_\diamondsuit(\theta^{*,\diamondsuit})$	$ASPD_\diamondsuit(\theta^{FA})$ $p=\frac{2}{3}$	$p=\frac{1}{2}$
(7, 4.5, 6, 3, 1)	(0.7, 0.3, 0.6, 0.4)	(0.5, 0.5)	0.2598	0.3037 $\uparrow$	0.3094 $\uparrow$
		(0.3, 0.7)	0.2589	0.2961 $\uparrow$	0.3120 $\uparrow$
		(0.7, 0.3)	0.2646	0.3098 $\uparrow$	0.3037 $\uparrow$
	(0.7, 0.3, 0.4, 0.6)	(0.5, 0.5)	0.4263	0.4079 $\downarrow$	0.3667 $\downarrow$
		(0.3, 0.7)	0.4288	0.4124 $\downarrow$	0.3765 $\downarrow$
		(0.7, 0.3)	0.4240	0.4013 $\downarrow$	0.3550 $\downarrow$
(7, 4, 6, 3.5, 1)	(0.7, 0.3, 0.6, 0.4)	(0.5, 0.5)	0.1871	0.2860 $\uparrow$	0.3026 $\uparrow$
		(0.3, 0.7)	0.1785	0.2795 $\uparrow$	0.3035 $\uparrow$
		(0.7, 0.3)	0.1984	0.2925 $\uparrow$	0.3006 $\uparrow$
	(0.7, 0.3, 0.4, 0.6)	(0.5, 0.5)	0.3576	0.3916 $\uparrow$	0.3595 $\uparrow$
		(0.3, 0.7)	0.3538	0.3922 $\uparrow$	0.3627 $\uparrow$
		(0.7, 0.3)	0.3620	0.3905 $\uparrow$	0.3554 $\downarrow$

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 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 12, 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. As we described earlier, it is clearly that $p = 1/2$ is not in the range of $(p_{low}^\diamondsuit, p_{high}^\diamondsuit)$.