

A Fairness Criteria in Centralized and Federated Learning Setting

In this section, we provide supplementary discussion of the fairness criteria and their corresponding confusion-matrix formulations under both centralized and federated learning settings. First, in addition to the demographic parity (DP) and equal opportunity (EOP) notions introduced above, we here present the definitions of equality of odds (EO) along with their confusion-matrix representations. Next, we clarify how these fairness notions are formalized within FL, specifying the distinct fairness metrics employed at both the global and the local levels. Note that this paper adopts a subgroup-like fairness metric [73, 13, 40] to reduce the number of constraints, while our confusion-matrix representation is also applicable to the group-wise definitions of these fairness metrics [70, 71].

A.1 Group Fairness Criteria

Probabilistic notations. We elucidate some probability notations in the Preliminaries 3 and Table 1. Here, we use p_δ to denote the probability of event δ occurring. For example, $p_a := \mathbb{P}(A = a)$, $p_y = \mathbb{P}(Y = y)$, $p_{a,k} := \mathbb{P}(A = a, K = k)$, $p_{k|a} := \mathbb{P}(K = k | A = a)$, $p_{a|k} := \mathbb{P}(A = a | K = k)$, $p_{a,y} := \mathbb{P}(A = a, Y = y)$, $p_{y,k} := \mathbb{P}(Y = y, K = k)$, and $p_{a,y,k} := \mathbb{P}(A = a, Y = y, K = k)$.

Confusion-matrix-based fairness notations. For random tuple (X, Y, A) , the prediction of the (attribute-aware) classifier is defined as $\hat{Y} = h(X, A)$. One may simply choose $\hat{Y} = h(X)$ to consider the attribute-blind setting. To represent group fairness constraints, previous works [73, 55] introduce the group-specific confusion matrices \mathbf{C}^a , $a \in \mathcal{A}$ to characterize the fairness constraints, where $\mathbf{C}_{i,j}^a := \mathbb{P}(Y = i, \hat{Y} = j | A = a)$.

Example 1. For DP criterion,

$$\mathcal{D}_{DP} = \max_{y \in [m]} \max_{a' \in \mathcal{A}} \left| \mathbb{P}(\hat{Y} = y | A = a') - \mathbb{P}(\hat{Y} = y) \right|,$$

where $\mathbb{P}(Y = y | A = a') = \sum_{i \in [m]} \mathbb{P}(\hat{Y} = y, Y = i | A = a') = \sum_{i \in [m]} \mathbf{C}_{i,y}^{a'}$ and $\mathbb{P}(\hat{Y} = y) = \sum_{a \in \mathcal{A}} \mathbb{P}(A = a) \sum_{i \in [m]} \mathbb{P}(\hat{Y} = y, Y = i | A = a) = \sum_{a \in \mathcal{A}} \sum_{i \in [m]} \mathbb{P}(A = a) \mathbf{C}_{i,y}^a$.

Hence, we have

$$\mathcal{D}_{DP} = \max_{y \in [m]} \max_{a' \in \mathcal{A}} \left| \sum_{a \in \mathcal{A}} \sum_{i \in [m]} (\mathbb{I}[a = a'] - \mathbb{P}(A = a)) \mathbf{C}_{i,y}^a \right| = \max_{y \in [m]} \max_{a' \in \mathcal{A}} \left| \sum_{a \in \mathcal{A}} \langle \mathbf{D}_{a',y}^a, \mathbf{C}^a \rangle \right|,$$

where $\mathbf{D}_{a',y}^a \in \mathbb{R}^{m \times m}$, and the y -th column elements of $\mathbf{D}_{a',y}^a$ are $\mathbb{I}[a = a'] - \mathbb{P}(A = a)$ with all other elements set to 0.

Example 2. For EOP criterion,

$$\mathcal{D}_{EOP} = \max_{y \in [m]} \max_{a' \in \mathcal{A}} \left| \mathbb{P}(\hat{Y} = y | A = a', Y = y) - \mathbb{P}(\hat{Y} = y | Y = y) \right|,$$

where $\mathbb{P}(Y = y | A = a', Y = y) = \frac{p_{a',y}}{p_{a',y}} \mathbf{C}_{y,y}^{a'}$ and $\mathbb{P}(\hat{Y} = y | Y = y) = \sum_{a \in \mathcal{A}} \frac{p_a}{p_y} \mathbf{C}_{y,y}^a$.

Hence, we have

$$\mathcal{D}_{EOP} = \max_{y \in [m]} \max_{a' \in \mathcal{A}} \left| \sum_{a \in \mathcal{A}} \left(\frac{p_{a'}}{p_{a',y}} \mathbb{I}[a = a'] - \frac{p_a}{p_y} \right) \mathbf{C}_{y,y}^a \right| = \max_{y \in [m]} \max_{a' \in \mathcal{A}} \left| \sum_{a \in \mathcal{A}} \langle \mathbf{D}_{a',y}^a, \mathbf{C}^a \rangle \right|,$$

where $\mathbf{D}_{a',y}^a \in \mathbb{R}^{m \times m}$, and the entry in the y -th row and y -th column is $\frac{p_{a'}}{p_{a',y}} \mathbb{I}[a = a'] - \frac{p_a}{p_y}$ with all other elements set to 0.

Example 3. For EO criterion, we follow [3] to introduce the mean equalized odds (MEO) constraint, and consider its subgroup-like representation:

$$\mathcal{D}_{EO} = \max_{y \in [m]} \max_{a' \in \mathcal{A}} \frac{1}{2} (|\text{TPR}_y(a) - \text{TPR}_y| + |\text{FPR}_y(a) - \text{FPR}_y|),$$

where $\text{TPR}_y(a) = \mathbb{P}(\hat{Y} = y | Y = y, A = a)$, $\text{TPR}_y = \mathbb{P}(\hat{Y} = y | Y = y)$ and $\text{FPR}_y(a) = \mathbb{P}(\hat{Y} = y | Y \neq y, A = a)$, $\text{FPR}_y = \mathbb{P}(\hat{Y} = y | Y \neq y)$.

934 It shows that

$$\begin{aligned}
& \frac{1}{2}(|\text{TPR}_y(a) - \text{TPR}_y| + |\text{FPR}_y(a) - \text{FPR}_y|) \\
&= \frac{1}{2} \left(\left| \sum_{a \in \mathcal{A}} \left(\frac{p_{a'}}{p_{a',y}} \mathbb{I}[a = a'] - \frac{p_a}{p_y} \right) \mathbf{C}_{y,y}^a \right| + \left| \sum_{a \in \mathcal{A}} \sum_{y_i \neq y} \left(\frac{p_{a'}}{\sum_{y_j \neq y} p_{a',y_j}} \mathbb{I}[a = a'] - \frac{p_a}{\sum_{y_j \neq y} p_{y_j}} \right) \mathbf{C}_{y_i,y}^a \right| \right), \\
&= \frac{1}{2} \left(\left| \sum_{a \in \mathcal{A}} \langle \mathbf{D}_{a',y}^{a,0}, \mathbf{C}^a \rangle \right| + \left| \sum_{a \in \mathcal{A}} \langle \mathbf{D}_{a',y}^{a,1}, \mathbf{C}^a \rangle \right| \right),
\end{aligned}$$

935 where the entry in the y -th row and y -th column is $\frac{p_{a'}}{p_{a',y}} \mathbb{I}[a = a'] - \frac{p_a}{p_y}$ with all other elements set to
936 0 for $\mathbf{D}_{a',y}^{a,0} \in \mathbb{R}^{m \times m}$, and the entry in the y -th column is $\frac{p_{a'}}{\sum_{y_j \neq y} p_{a',y_j}} \mathbb{I}[a = a'] - \frac{p_a}{\sum_{y_j \neq y} p_{y_j}}$ except
937 for the y -th row with all other elements set to 0 for $\mathbf{D}_{a',y}^{a,1} \in \mathbb{R}^{m \times m}$.

938 A.2 Group Fairness notations in FL

939 As noted in the main text, fairness at the level of each client's dataset (*local fairness*) differs from
940 fairness across the aggregate dataset of all clients (*global fairness*). Local fairness is defined with
941 respect to each client's individual data distribution $\mathbb{P}(X, Y, A \mid K)$, whereas global fairness is
942 defined over the overall (aggregate) distribution $\mathbb{P}(X, Y, A)$. Motivated by approaches that employ
943 group-specific confusion matrices for fairness [73, 55], we propose the **decentralized group-specific**
944 **confusion matrices** $\mathbf{C}^{a,k}$, $a \in \mathcal{A}, k \in [N]$ to capture both global and local fairness across multiple
945 data distributions within FL, with elements defined for $i, j \in [m]$ as $\mathbf{C}_{i,j}^{a,k}(h) := \mathbb{P}(Y = i, \hat{Y} = j \mid$
946 $A = a, K = k)$.

947 **Example 4.** For DP criterion, the global DP fairness metric is defined as

$$\mathcal{D}_{DP}^g = \max_{y \in [m]} \max_{a' \in \mathcal{A}} \left| \mathbb{P}(\hat{Y} = y \mid A = a') - \mathbb{P}(\hat{Y} = y) \right|,$$

948 where $\mathbb{P}(Y = y \mid A = a') = \sum_{k \in [N]} \sum_{i \in [m]} p_{k|a'} \mathbf{C}_{i,y}^{a',k}(h_k)$, and $\mathbb{P}(\hat{Y} = y) =$
949 $\sum_{a \in \mathcal{A}} \sum_{k \in [N]} \sum_{i \in [m]} p_{a,k} \mathbf{C}_{i,y}^{a,k}(h_k)$. Hence, we have

$$\begin{aligned}
\mathcal{D}_{DP}^g &= \max_{y \in [m]} \max_{a' \in \mathcal{A}} \left| \sum_{a \in \mathcal{A}} \sum_{k \in [N]} \sum_{i \in [m]} (p_{k|a'} \mathbb{I}[a = a'] - p_{a,k}) \mathbf{C}_{i,y}^{a,k}(h_k) \right| \\
&= \max_{y \in [m]} \max_{a' \in \mathcal{A}} \left| \sum_{a \in \mathcal{A}} \sum_{k \in [N]} \langle \mathbf{D}_{a',y}^{a,k}, \mathbf{C}^{a,k}(h_k) \rangle \right|,
\end{aligned}$$

950 where $\mathbf{D}_{a',y}^{a,k} \in \mathbb{R}^{m \times m}$, and the y -th column elements of $\mathbf{D}_{a',y}^{a,k}$ are $\mathbb{P}(K = k \mid A = a') \mathbb{I}[a =$
951 $a'] - \mathbb{P}(A = a, K = k)$ with all other elements set to 0.

952 The local DP fairness metric for k -th client is defined as

$$\mathcal{D}_{DP}^{l,k} = \max_{y \in [m]} \max_{a' \in \mathcal{A}} \left| \mathbb{P}(\hat{Y} = y \mid A = a', K = k) - \mathbb{P}(\hat{Y} = y \mid K = k) \right|,$$

953 where $\mathbb{P}(Y = y \mid A = a', K = k) = \sum_{i \in [m]} \mathbf{C}_{i,y}^{a',k}$, and $\mathbb{P}(\hat{Y} = y \mid K = k) =$
954 $\sum_{a \in \mathcal{A}} \sum_{i \in [m]} p_{a|k} \mathbb{P}(\hat{Y} = y, Y = i \mid A = a, K = k)$. Hence, we have

$$\mathcal{D}_{DP}^{l,k} = \max_{y \in [m]} \max_{a' \in \mathcal{A}} \left| \sum_{a \in \mathcal{A}} \sum_{i \in [m]} (\mathbb{I}[a = a'] - p_{a|k}) \mathbf{C}_{i,y}^{a,k}(h_k) \right| = \max_{y \in [m]} \max_{a' \in \mathcal{A}} \left| \sum_{a \in \mathcal{A}} \langle \mathbf{D}_{a',y}^{a,k}, \mathbf{C}^{a,k}(h_k) \rangle \right|,$$

955 where $\mathbf{D}_{a',y}^{a,k} \in \mathbb{R}^{m \times m}$, and the y -th column elements of $\mathbf{D}_{a',y}^{a,k}$ are $\mathbb{I}[a = a'] - \mathbb{P}(A = a \mid K = k)$
956 with all other elements set to 0.

957 **Example 5.** For EOP criterion, the global EOP fairness metric is defined as

$$\mathcal{D}_{EOP}^g = \max_{y \in [m]} \max_{a' \in \mathcal{A}} \left| \mathbb{P}(\hat{Y} = y \mid Y = y, A = a') - \mathbb{P}(\hat{Y} = y \mid Y = y) \right|,$$

958 where $\mathbb{P}(Y = y \mid Y = y, A = a') = \sum_{k \in [N]} \frac{p_{a',k}}{p_{a',y}} \mathbf{C}_{i,y}^{a',k}(h_k)$, and $\mathbb{P}(\hat{Y} = y \mid Y = y) =$
 959 $\sum_{a \in \mathcal{A}} \sum_{k \in [N]} \frac{p_{a,k}}{p_y} \mathbf{C}_{i,y}^{a,k}(h_k)$. Hence, we have

$$\begin{aligned} \mathcal{D}_{DP}^g &= \max_{y \in [m]} \max_{a' \in \mathcal{A}} \left| \sum_{a \in \mathcal{A}} \sum_{k \in [N]} \left(\frac{p_{a',k}}{p_{a',y}} \mathbb{I}[a = a'] - \frac{p_{a,k}}{p_y} \right) \mathbf{C}_{i,y}^{a,k}(h_k) \right| \\ &= \max_{y \in [m]} \max_{a' \in \mathcal{A}} \left| \sum_{a \in \mathcal{A}} \sum_{k \in [N]} \langle \mathbf{D}_{a',y}^{a,k}, \mathbf{C}_{i,y}^{a,k}(h_k) \rangle \right|, \end{aligned}$$

960 where $\mathbf{D}_{a',y}^{a,k} \in \mathbb{R}^{m \times m}$, and the entry in the y -th row and y -th column is $\frac{p_{a',k}}{p_{a',y}} \mathbb{I}[a = a'] - \frac{p_{a,k}}{p_y}$ with
 961 all other elements set to 0.

962 The local EOP fairness metric for k -th client is defined as

$$\mathcal{D}_{EOP}^{l,k} = \max_{y \in [m]} \max_{a' \in \mathcal{A}} \left| \mathbb{P}(\hat{Y} = y \mid A = a', Y = y, K = k) - \mathbb{P}(\hat{Y} = y \mid Y = y, K = k) \right|,$$

963 where $\mathbb{P}(\hat{Y} = y \mid A = a', Y = y, K = k) = \frac{p_{a',k}}{p_{a',y,k}} \mathbf{C}_{i,y}^{a',k}$, and $\mathbb{P}(\hat{Y} = y \mid Y = y, K = k) =$
 964 $\sum_{a \in \mathcal{A}} \frac{p_{a,k}}{p_{y,k}} \mathbb{P}(\hat{Y} = y, Y = i \mid A = a, K = k)$. Hence, we have

$$\mathcal{D}_{EOP}^{l,k} = \max_{y \in [m]} \max_{a' \in \mathcal{A}} \left| \sum_{a \in \mathcal{A}} \left(\frac{p_{a',k}}{p_{a',y,k}} \mathbb{I}[a = a'] - \frac{p_{a,k}}{p_{y,k}} \right) \mathbf{C}_{i,y}^{a,k}(h_k) \right| = \max_{y \in [m]} \max_{a' \in \mathcal{A}} \left| \sum_{a \in \mathcal{A}} \langle \mathbf{D}_{a',y}^{a,k}, \mathbf{C}_{i,y}^{a,k}(h_k) \rangle \right|,$$

965 where $\mathbf{D}_{a',y}^{a,k} \in \mathbb{R}^{m \times m}$, and the entry in the y -th row and y -th column is $\frac{p_{a',k}}{p_{a',y,k}} \mathbb{I}[a = a'] - \frac{p_{a,k}}{p_{y,k}}$ with
 966 all other elements set to 0.

967 **Example 6.** For EO criterion, the global EO fairness metric is defined as

$$\begin{aligned} \mathcal{D}_{EO}^g &= \max_{y \in [m]} \max_{a' \in \mathcal{A}} \frac{1}{2} \left(\left| \mathbb{P}(\hat{Y} = y \mid Y = y, A = a') - \mathbb{P}(\hat{Y} = y \mid Y = y) \right| \right. \\ &\quad \left. + \left| \mathbb{P}(\hat{Y} = y \mid Y \neq y, A = a') - \mathbb{P}(\hat{Y} = y \mid Y \neq y) \right| \right), \end{aligned}$$

968 where $\mathbb{P}(Y = y \mid Y \neq y, A = a') = \sum_{k \in [N]} \sum_{y_i \neq y} \frac{p_{a',k}}{\sum_{y_j \neq y} p_{a',y_j}} \mathbf{C}_{y_i,y}^{a',k}(h_k)$, and $\mathbb{P}(\hat{Y} = y \mid Y \neq$
 969 $y) = \sum_{a \in \mathcal{A}} \sum_{k \in [N]} \sum_{y_i \neq y} \frac{p_{a,k}}{\sum_{y_j \neq y} p_{y_j}} \mathbf{C}_{y_i,y}^{a,k}(h_k)$. Hence, we have

$$\mathcal{D}_{EO}^g = \max_{y \in [m]} \max_{a' \in \mathcal{A}} \frac{1}{2} \left(\left| \sum_{a \in \mathcal{A}} \sum_{k \in [N]} \langle \mathbf{D}_{a',y}^{a,k,0}, \mathbf{C}_{i,y}^{a,k}(h_k) \rangle \right| + \left| \sum_{a \in \mathcal{A}} \sum_{k \in [N]} \langle \mathbf{D}_{a',y}^{a,k,1}, \mathbf{C}_{i,y}^{a,k}(h_k) \rangle \right| \right),$$

970 where the entry in the y -th row and y -th column is $\frac{p_{a',k}}{p_{a',y}} \mathbb{I}[a = a'] - \frac{p_{a,k}}{p_y}$ with all other elements set to
 971 0 for $\mathbf{D}_{a',y}^{a,k,0} \in \mathbb{R}^{m \times m}$, and the entry in the y -th column is $\frac{p_{a',k}}{\sum_{y_j \neq y} p_{a',y_j}} \mathbb{I}[a = a'] - \frac{p_{a,k}}{\sum_{y_j \neq y} p_{y_j}}$ except
 972 for the y -th row with all other elements set to 0 for $\mathbf{D}_{a',y}^{a,k,1} \in \mathbb{R}^{m \times m}$.

973 The local EO fairness metric for k -th client is defined as

$$\begin{aligned} \mathcal{D}_{EO}^{l,k} &= \max_{y \in [m]} \max_{a' \in \mathcal{A}} \frac{1}{2} \left(\left| \mathbb{P}(\hat{Y} = y \mid Y = y, A = a', K = k) - \mathbb{P}(\hat{Y} = y \mid Y = y, K = k) \right| \right. \\ &\quad \left. + \left| \mathbb{P}(\hat{Y} = y \mid Y \neq y, A = a', K = k) - \mathbb{P}(\hat{Y} = y \mid Y \neq y, K = k) \right| \right), \end{aligned}$$

974 where $\mathbb{P}(\hat{Y} = y \mid A = a', Y \neq y, K = k) = \sum_{y_i \neq y} \frac{p_{a',k}}{\sum_{y_j \neq y} p_{a',y_j,k}} \mathbf{C}_{y_i,y}^{p_{a',k}}$, and $\mathbb{P}(\hat{Y} = y \mid Y \neq$
975 $y, K = k) = \sum_{a \in \mathcal{A}} \frac{p_{a,k}}{\sum_{y_j \neq y} p_{y_j,k}} \mathbb{P}(\hat{Y} = y, Y = i \mid A = a, K = k)$. Hence, we have

$$\mathcal{D}_{EO}^g = \max_{y \in [m]} \max_{a' \in \mathcal{A}} \frac{1}{2} \left(\left| \sum_{a \in \mathcal{A}} \sum_{k \in [N]} \langle \mathbf{D}_{a',y}^{a,k,0}, \mathbf{C}^{a,k}(h_k) \rangle \right| + \left| \sum_{a \in \mathcal{A}} \sum_{k \in [N]} \langle \mathbf{D}_{a',y}^{a,k,1}, \mathbf{C}^{a,k}(h_k) \rangle \right| \right),$$

976 where the entry in the y -th row and y -th column is $\frac{a',k}{p_{a',y,k}} \mathbb{I}[a = a'] - \frac{p_{a,k}}{p_{y,k}}$ with all other elements set
977 to 0 for $\mathbf{D}_{a',y}^{a,k,0} \in \mathbb{R}^{m \times m}$, and the entry in the y -th column is $\frac{p_{a',k}}{\sum_{y_j \neq y} p_{a',y_j,k}} \mathbb{I}[a = a'] - \frac{p_{a,k}}{\sum_{y_j \neq y} p_{y_j,k}}$
978 except for the y -th row with all other elements set to 0 for $\mathbf{D}_{a',y}^{a,k,1} \in \mathbb{R}^{m \times m}$.

B Proofs and Discussion in Section 4

B.1 Proof of Proposition 1

This section provides the proof of Proposition 1. The proof is primarily inspired by the characterization of the Bayes-optimal fair classifier in the centralized fair machine learning literature (e.g. Theorem 3.1 of [73], Proposition 10 of [55]).

Proof. We begin by casting the primal problem (1) into an optimization problem defined on the Cartesian product of confusion matrices. Consider the the set of achievable confusion matrices:

$$\mathcal{C}^{|\mathcal{A}|\times N} := \{\mathbf{C}^{|\mathcal{A}|\times N}(\mathbf{h}) := \{\mathbf{C}^{a,k}(h_k)\}_{a\in\mathcal{A}, k\in[N]} : \mathbf{h} \in \mathcal{H}^N\},$$

where $\mathcal{C}^{|\mathcal{A}|\times N}$ be the product space of all confusion matrices $\mathbf{C}^{a,k}$ corresponding to sensitive group $a \in \mathcal{A}$ and $k \in [N]$ associated with a given instance $\mathbf{h} \in \mathcal{H}^N$ of the problem. It is clear that the performance metric \mathcal{R} and fairness metrics $\mathcal{D}^g, \mathcal{D}^{l,k}, k \in [N]$ are continuous and bounded to $\mathcal{C}^{|\mathcal{A}|\times N}(\mathbf{h}) := \{\mathbf{C}^{a,k}(h_k)\}_{a\in\mathcal{A}, k\in[N]}$.

Convexity of $\mathcal{C}^{|\mathcal{A}|\times N}$. Let $\forall \mathbf{C}_1, \mathbf{C}_2 \in \mathcal{C}^{|\mathcal{A}|\times N}$ be realized by classifier tuples $\mathbf{h}_1, \mathbf{h}_2$. For any $\omega \in [0, 1]$, define the mixed classifier $\mathbf{h}' = \omega \mathbf{h}_1 + (1 - \omega) \mathbf{h}_2$. By linearity of performance and fairness metrics, its confusion matrix satisfies

$$\mathbf{C}(\mathbf{h}') = \omega \mathbf{C}(\mathbf{h}_1) + (1 - \omega) \mathbf{C}(\mathbf{h}_2) = \omega \mathbf{C}_1 + (1 - \omega) \mathbf{C}_2 = \mathbf{C}_\omega.$$

Thus every convex combination of \mathbf{C}_1 and \mathbf{C}_2 lies in $\mathcal{C}^{|\mathcal{A}|\times N}$, establishing convexity.

Deterministic classifiers. It can be seen that, for any linear objective $\phi_{\mathbf{L}}(\mathbf{C}^{|\mathcal{A}|\times N}(\mathbf{h})) = \sum_{a\in\mathcal{A}} \sum_{k\in[N]} \langle \mathbf{L}^{a,k}, \mathbf{C}^{a,k}(\mathbf{h}_k) \rangle$, there is a deterministic classifiers $\mathbf{h}^* = (h_1^*, \dots, h_N^*)$ that is optimal for $\phi_{\mathbf{L}}$ (see proof in B.2). By the supporting-hyperplane theorem [8] for compact convex sets, for each point $\mathbf{C}_b = \{\mathbf{C}_b^{a,k}\}_{a\in\mathcal{A}, k\in[N]} \in \partial \mathcal{C}^{|\mathcal{A}|\times N}$, there exists a nonzero collection of matrices $\mathbf{L}_b = \{\mathbf{L}_b^{a,k}\}_{a\in\mathcal{A}, k\in[N]}$ constitutes a hyperplane, such that for every $\mathbf{C} = \{\mathbf{C}^{a,k}\} \in \mathcal{C}^{|\mathcal{A}|\times N}$ we have $\sum_{a\in\mathcal{A}} \sum_{k=1}^N \langle \mathbf{L}_b^{a,k}, \mathbf{C}_b^{a,k} \rangle \leq \sum_{a\in\mathcal{A}} \sum_{k=1}^N \langle \mathbf{L}_b^{a,k}, \mathbf{C}^{a,k} \rangle$ which is precisely the desired supporting-hyperplane condition at \mathbf{C}_b . In other words, we arrive at the conclusion that each boundary point of $\mathcal{C}^{|\mathcal{A}|\times N}$ can be achieved by deterministic classifiers $\mathbf{h}' = (h'_1, \dots, h'_N)$.

Combination of deterministic classifiers. Since $\mathcal{C}^{|\mathcal{A}|\times N}$ is compact and convex, we know that its extreme points fall in its boundary. By the Krein-Milman theorem [56], we have that $\mathcal{C}^{|\mathcal{A}|\times N}$ is equal to the convex hull of its extreme points. We further have from Caratheodory's theorem [8] that any $\mathbf{C} \in \mathcal{C}^{|\mathcal{A}|\times N}$ can be expressed as a convex combination of $d_k = |\mathcal{A}|Nm^2$ points in the extreme point set, where each extreme point can be characterized by deterministic classifiers. Hence, we have proved that the optimal solution \mathbf{h} can be represented by the convex combination of deterministic classifiers. \square

Discussion on feasibility. The only condition for the above theorem to hold is that the feasible set is non-empty, which is clearly satisfied by the mentioned fairness constraints, DP, EOP, and EO. For these fairness criteria, the classifier that always predicts a single, fixed label y' trivially meets $\xi^g = 0, \xi^{l,k} = 0, k \in [N]$, and hence satisfies the fairness constraints.

The number of deterministic classifiers. As for the number of deterministic classifiers required, the parameter d^k in the proof scales with the number of nonzero entries in the linear performance and fairness constraints [55]. Since each matrix $\mathbf{D}^{a,k}$ in our fairness formulation is zero except for one column, we in fact need far fewer than $|\mathcal{A}|Nm^2$ classifiers. Moreover, under the continuity assumption 1, this number can be reduced even further [73].

B.2 Proof of Proposition 2

Proof. We denote $p_a := \mathbb{P}(A = a)$, $p_k := \mathbb{P}(K = k)$, $p_{a,k} := \mathbb{P}(A = a, K = k)$, and $\mathcal{P}_k^X := \mathbb{P}(X|K = k)$. Consider the form Lagrangian function of federated Bayes-optimal fair classification

1021 problem (1),

$\mathcal{L}(\mathbf{h}, \lambda, \mu)$

$$\begin{aligned}
&= \sum_{k=1}^N \sum_{a \in \mathcal{A}} p_{a,k} \langle \mathbf{1} - \mathbf{I}, \mathbf{C}^{a,k}(h_k) \rangle + \sum_{u_g \in \mathcal{U}_g} (\lambda_{u_g}^{(1)} - \lambda_{u_g}^{(2)}) \sum_{k=1}^N \sum_{a \in \mathcal{A}} \langle \mathbf{D}_{u_g}^{a,k}, \mathbf{C}^{a,k}(h_k) \rangle \\
&\quad + \sum_{k=1}^N \sum_{u_l, k \in \mathcal{U}_{l,k}} (\mu_{k,u_l,k}^{(1)} - \mu_{k,u_l,k}^{(2)}) \sum_{a \in \mathcal{A}} \langle \mathbf{D}_{u_l,k}^{a,k}, \mathbf{C}^{a,k}(h_k) \rangle - \sum_{u_g \in \mathcal{U}_g} (\lambda_{u_g}^{(1)} + \lambda_{u_g}^{(2)}) \xi^g \\
&\quad - \sum_{k \in [N]} \sum_{u_l, k \in \mathcal{U}_{l,k}} (\mu_{k,u_l,k}^{(1)} + \mu_{k,u_l,k}^{(2)}) \xi^{l,k} \\
&= \sum_{k=1}^N \sum_{a \in \mathcal{A}} \left\langle p_{a,k} (\mathbf{1} - \mathbf{I}) + \sum_{u_g \in \mathcal{U}_g} (\lambda_{u_g}^{(1)} - \lambda_{u_g}^{(2)}) \mathbf{D}_{u_g}^{a,k} + \sum_{u_l, k \in \mathcal{U}_{l,k}} (\mu_{k,u_l,k}^{(1)} - \mu_{k,u_l,k}^{(2)}) \mathbf{D}_{u_l,k}^{a,k}, \mathbf{C}^{a,k}(h_k) \right\rangle \\
&\quad - \sum_{u_g \in \mathcal{U}_g} (\lambda_{u_g}^{(1)} + \lambda_{u_g}^{(2)}) \xi^g - \sum_{k \in [N]} \sum_{u_l, k \in \mathcal{U}_{l,k}} (\mu_{k,u_l,k}^{(1)} + \mu_{k,u_l,k}^{(2)}) \xi^{l,k}.
\end{aligned}$$

1022 The inner problem of Lagrangian dual ask we to solve $\min_{\mathbf{h} \in \mathcal{H}} \mathcal{L}(\mathbf{h}, \lambda, \mu)$ given element-wise
1023 non-negative dual parameter λ and μ , which can be formulated as

$$\max_{\mathbf{h} \in \mathcal{H}^N} V(\mathbf{h}, \lambda, \mu) = \sum_{k=1}^N \sum_{a \in \mathcal{A}} p_{a,k} \langle \mathbf{M}^{\lambda, \mu}(a, k), \mathbf{C}^{a,k}(h_k) \rangle,$$

1024 where $\mathbf{M}^{\lambda, \mu}(a, k) := \mathbf{I} - \frac{1}{p_{a,k}} \left[\sum_{u_g \in \mathcal{U}_g} (\lambda_{u_g}^{(1)} - \lambda_{u_g}^{(2)}) \mathbf{D}_{u_g}^{a,k} - \sum_{u_l, k \in \mathcal{U}_{l,k}} (\mu_{k,u_l,k}^{(1)} - \mu_{k,u_l,k}^{(2)}) \mathbf{D}_{u_l,k}^{a,k} \right]$.

1025 The next step is to derive the optimal solution of $\max_{\mathbf{h} \in \mathcal{H}^N} V(\mathbf{h}, \lambda, \mu)$. For this purpose, we perform
1026 manipulations of H to reveal its clear relationship with the personalized classifier $\mathbf{h} = (h_1, \dots, h_N)$.
1027 Denote the condition distribution of X given sensitive attribute $A = a$ on client $K = k$ as $\mathcal{P}_{a,k}^X$, i.e.,

1028 $\mathcal{P}_{a,k}^X := \mathbb{P}(X | A = a, K = k)$, we have

$$\begin{aligned}
V(\mathbf{h}, \lambda, \mu) &= \sum_{k=1}^N \sum_{a \in \mathcal{A}} p_{a,k} \langle \mathbf{M}^{\lambda, \mu}(a, k), \mathbf{C}^{a,k}(h_k) \rangle \\
&= \sum_{k=1}^N \sum_{a \in \mathcal{A}} p_{a,k} \int_{\mathcal{X}} [\eta(X, a, k)]^\top \mathbf{M}^{\lambda, \mu}(a, k) h_k(x) d\mathcal{P}_{a,k}^X(x) \\
&= \sum_{k=1}^N \sum_{a \in \mathcal{A}} p_{a,k} \mathbb{E}_{X|A=a, K=k} \left[[\eta(X, a, k)]^\top \mathbf{M}^{\lambda, \mu}(a, k) h_k(x) \right] \\
&= \mathbb{E}_{A,K} \left[\mathbb{E}_{X|A,K} \left[[\eta(X, A, K)]^\top \mathbf{M}^{\lambda, \mu}(A, K) h_K(X) \right] \right] \\
&= \mathbb{E}_{X,A,K} \left[[\eta(X, A, K)]^\top \mathbf{M}^{\lambda, \mu}(A, K) h_K(X) \right] \\
&= \mathbb{E}_{X,K} \left[\mathbb{E}_{A|X,K} \left[[\eta(X, A, K)]^\top \mathbf{M}^{\lambda, \mu}(A, K) h_K(X) \right] \right] \\
&= \mathbb{E}_{X,K} \left[\sum_{a \in \mathcal{A}} \mathbb{P}(A = a | X, K) [\eta(X, a, K)]^\top \mathbf{M}^{\lambda, \mu}(A, K) h_K(X) \right] \\
&= \sum_{k=1}^N p_k \mathbb{E}_{x \sim \mathcal{P}_k^X} \left[\sum_{a \in \mathcal{A}} \mathbb{P}(A = a | x, k) [\eta(x, a, k)]^\top \mathbf{M}^{\lambda, \mu}(a, k) h_k(x) \right].
\end{aligned}$$

1029 To derive the optimal solution of the inner optimization problem, it suffices to perform a pointwise
1030 maximization of the above objective: for fixed x, k , the classifier $h_k(x)$ selects the label that
1031 maximizes the term inside the expectation, i.e.,

$$h_k^*(x) = e_y, \quad y \in \arg \max_{j \in [m]} \left(\sum_{a \in \mathcal{A}} \mathbb{P}(A = a | x, k) [\mathbf{M}^{\mu, \lambda}(a, k)]^\top \eta(x, a, k) \right)_j.$$

1032 Thus, we have finished the proof of Proposition 2. \square

1033 B.3 Proof of Proposition 3 and Further Exploration

1034 In this section, we prove that the representation in (4) is calibrated for both unified and personalized
1035 inner optimization problem. We begin by presenting the following lemma.

Lemma B.1. *For any categorical distribution characterized by $\mathbf{p} \in \Delta_m$, the minimizer of the expected risk*

$$\mathbb{E}_{y \sim \mathbf{p}} [-\log(\mathbf{q}_y)] = - \sum_{i=1}^m \mathbf{p}_i \log(\mathbf{q}_i)$$

1036 over all $\mathbf{q} \in \Delta_m$ is unique and achieved at $\mathbf{p} = \mathbf{q}$.

1037 This lemma is commonly used in the design of multiclass loss functions [69, 52, 54].

1038 B.3.1 Proof of Proposition 3

1039 *Proof.* We aim to prove that for any fixed $x \in \mathcal{X}, k \in [N]$, the optimal personalized scoring
1040 function $\mathbf{s}_k^* : \mathcal{X} \rightarrow \mathbb{R}^m$ that minimizes the expected loss $\ell_k(y, \mathbf{s}(x), a)$ over the local data distri-
1041 bution $\mathbb{P}(X, A, Y \mid K = k)$ recovers the personalized federated Bayes-optimal classifier $h_k^*(x)$ in
1042 Proposition 2.

1043 It is equivalent to show that, for any x :

$$\arg \max_{j \in [m]} [\mathbf{s}_k^*(x)]_j \subseteq \arg \max_{j \in [m]} \left(\sum_{a \in \mathcal{A}} P(A = a | x, k) [\mathbf{M}^{\mu, \lambda}(a, k)]^\top \eta(x, a, k) \right)_j.$$

1044 To this end, by leveraging the properties of conditional expectation, the cost-sensitive loss is reformu-
1045 lated as a function of the marginal distribution (X, K) :

$$\begin{aligned} \mathbb{E}_{(x, y, a, k) \sim (X, Y, A, K)} [\ell_k(y, \mathbf{s}(x), a)] &= -\mathbb{E}_{X, Y, A, K} \left[\sum_{i=1}^m \overline{\mathbf{M}}_{Y, i}^{\lambda, \mu}(A, K) \log \frac{\exp([\mathbf{s}(X)]_i)}{\sum_{j=1}^m \exp([\mathbf{s}(X)]_j)} \right] \\ &= -\mathbb{E}_{X, A, K} \left[\mathbb{E}_{Y | X, A, K} \left[\sum_{i=1}^m \overline{\mathbf{M}}_{Y, i}^{\lambda, \mu}(A, K) \log \frac{\exp([\mathbf{s}(X)]_i)}{\sum_{j=1}^m \exp([\mathbf{s}(X)]_j)} \right] \right] \\ &= -\mathbb{E}_{X, A, K} \left[\sum_{y \in [m]} \mathbb{P}(Y = y \mid X, A, K) \sum_{i=1}^m \overline{\mathbf{M}}_{y, i}^{\lambda, \mu}(A, K) \log \frac{\exp([\mathbf{s}(X)]_i)}{\sum_{j=1}^m \exp([\mathbf{s}(X)]_j)} \right] \\ &= -\mathbb{E}_{X, K} \left[\mathbb{E}_{A | X, K} \left[\sum_{y \in [m]} \eta_y(X, A, K) \sum_{i=1}^m \overline{\mathbf{M}}_{y, i}^{\mu, \lambda}(A, K) \log \frac{\exp([\mathbf{s}(X)]_i)}{\sum_{j=1}^m \exp([\mathbf{s}(X)]_j)} \right] \right] \\ &= \sum_{k=1}^N p_k \mathbb{E}_{x \sim \mathcal{P}_k^X} \left[- \sum_{a \in \mathcal{A}} \mathbb{P}(A = a | x, k) \sum_{i=1}^m \left([\overline{\mathbf{M}}^{\mu, \lambda}(a, k)]^\top \eta(x, a, k) \right)_i \log \frac{\exp([\mathbf{s}(x)]_i)}{\sum_{j=1}^m \exp([\mathbf{s}(x)]_j)} \right] \end{aligned}$$

1046 Denoting $\mathbf{v}_i(x, k) := \sum_{a \in \mathcal{A}} \mathbb{P}(A = a | x, k) \left([\overline{\mathbf{M}}^{\mu, \lambda}(a, k)]^\top \eta(x, a, k) \right)_i$, we have

$$\mathbb{E}_{X, Y, A, K} [\ell_k(y, \mathbf{s}(x), a)] = \mathbb{E}_{X, K} \left[-c_{X, K} \sum_{i=1}^m \frac{\mathbf{v}_i(X, K)}{\sum_{j \in [m]} \mathbf{v}_j(X, K)} \log \frac{\exp([\mathbf{s}(X)]_i)}{\sum_{j=1}^m \exp([\mathbf{s}(X)]_j)} \right]$$

1047 where $c_{x, k} = \sum_{j \in [m]} \mathbf{v}_j(x, k)$ can be treated as a constant for fixed x, k . According to Lemma B.1,
1048 given fixed x, k , an optimal personalized classifier $\mathbf{s}_k^*(x)$ minimizing the cost-sensitive loss point-wise
1049 satisfies

$$\frac{\mathbf{v}_i(x, k)}{\sum_{j \in [m]} \mathbf{v}_j(x, k)} = \frac{\exp([\mathbf{s}_k^*(x)]_i)}{\sum_{j=1}^m \exp([\mathbf{s}_k^*(x)]_j)}, \quad \forall i \in [m].$$

1050 It presents that, for all $i \in [m]$, since $\sum_{i \in [m]} \eta(x, a, k) = 1$,

$$\begin{aligned}
[\mathbf{s}_k^*(x)]_i &= \mathbf{v}_i(x, k) = \sum_{a \in \mathcal{A}} \mathbb{P}(A = a | x, k) \left([\bar{\mathbf{M}}^{\mu, \lambda}(a, k)]^\top \eta(x, a, k) \right)_i \\
&= \sum_{a \in \mathcal{A}} \mathbb{P}(A = a | x, k) \left([\mathbf{M}^{\mu, \lambda}(a, k) + \alpha \mathbf{1}_{m \times m}]^\top \eta(x, a, k) \right)_i \\
&= \sum_{a \in \mathcal{A}} \mathbb{P}(A = a | x, k) \left([\mathbf{M}^{\mu, \lambda}(a, k)]^\top \eta(x, a, k) + \alpha \mathbf{1}_m \right)_i \\
&= \sum_{a \in \mathcal{A}} \mathbb{P}(A = a | x, k) \left([\mathbf{M}^{\mu, \lambda}(a, k)]^\top \eta(x, a, k) \right)_i + \alpha.
\end{aligned}$$

1051 Hence,

$$\arg \max_{y \in [m]} [\mathbf{s}_k^*(x)]_y \subseteq \arg \max_{y \in [m]} \left(\sum_{a \in \mathcal{A}} \mathbb{P}(A = a | x, k) [\mathbf{M}^{\mu, \lambda}(a, k)]^\top \eta(x, a, k) \right)_y.$$

1052 The personalized classifier $h_k^*(x) \in \arg \min_{y \in [m]} [\mathbf{s}_k^*(x)]_y$ recovers that in Proposition 2. We finish
1053 the proof. \square

1054 B.3.2 Exploration of Calibrated Loss for Unified Bayes-Optimal Classifier

1055 We start from the inner optimization objective $V(\mathbf{h}, \lambda, \mu)$,

$$\begin{aligned}
V(\mathbf{h}, \lambda, \mu) &= \sum_{k=1}^N \sum_{a \in \mathcal{A}} p_{a,k} \left\langle \mathbf{M}^{\lambda, \mu}(a, k), \mathbf{C}^{a,k}(h_k) \right\rangle \\
&= \mathbb{E}_{X, A, K} \left[[\eta(X, A, K)]^\top \mathbf{M}^{\lambda, \mu}(A, K) h(X) \right] \\
&= \mathbb{E}_X \left[\mathbb{E}_{A, K | X} \left[[\eta(X, A, K)]^\top \mathbf{M}^{\lambda, \mu}(A, K) h(X) \right] \right] \\
&= \mathbb{E}_X \left[\sum_{a \in \mathcal{A}} \sum_{k=1}^N \mathbb{P}(A = a, K = k | X) [\eta(X, a, k)]^\top \mathbf{M}^{\lambda, \mu}(a, k) h(X) \right]
\end{aligned}$$

1056 To derive the optimal solution of the inner optimization problem, it suffices to perform a point-wise
1057 maximization of the above objective: for fixed x , the classifier $h(x)$ selects the label that maximizes
1058 the term inside the expectation, i.e.,

$$h^*(x) = e_y, \quad y \in \arg \max_{j \in [m]} \left(\sum_{a \in \mathcal{A}} \sum_{k=1}^N \mathbb{P}(A = a, K = k | X) [\eta(X, a, k)]^\top \mathbf{M}^{\lambda, \mu}(a, k) h(X) \right)_j.$$

1059 Consider the calibrated loss function in (4),

$$\begin{aligned}
\mathbb{E}_{(x,y,a,k) \sim (X,Y,A,K)}[\ell_k(y, \mathbf{s}(x), a)] &= -\mathbb{E}_{X,Y,A,K} \left[\sum_{i=1}^m \bar{\mathbf{M}}_{Y,i}^{\lambda,\mu}(A, K) \log \frac{\exp([\mathbf{s}(X)]_i)}{\sum_{j=1}^m \exp([\mathbf{s}(X)]_j)} \right] \\
&= -\mathbb{E}_{X,A,K} \left[\mathbb{E}_{Y|X,A,K} \left[\sum_{i=1}^m \bar{\mathbf{M}}_{Y,i}^{\lambda,\mu}(A, K) \log \frac{\exp([\mathbf{s}(X)]_i)}{\sum_{j=1}^m \exp([\mathbf{s}(X)]_j)} \right] \right] \\
&= -\mathbb{E}_{X,A,K} \left[\sum_{y \in [m]} \mathbb{P}(Y = y | X, A, K) \sum_{i=1}^m \bar{\mathbf{M}}_{y,i}^{\lambda,\mu}(A, K) \log \frac{\exp([\mathbf{s}(X)]_i)}{\sum_{j=1}^m \exp([\mathbf{s}(X)]_j)} \right] \\
&= -\mathbb{E}_X \left[\mathbb{E}_{A,K|X} \left[\sum_{y \in [m]} \eta_y(X, A, K) \sum_{i=1}^m \bar{\mathbf{M}}_{y,i}^{\mu,\lambda}(A, K) \log \frac{\exp([\mathbf{s}(X)]_i)}{\sum_{j=1}^m \exp([\mathbf{s}(X)]_j)} \right] \right] \\
&= \mathbb{E}_{x \sim \mathbb{P}(X)} \left[- \sum_{a \in \mathcal{A}} \sum_{k=1}^N \mathbb{P}(A = a, K = k | x) \sum_{i=1}^m \left(\left[\bar{\mathbf{M}}^{\mu,\lambda}(a, k) \right]^\top \eta(x, a, k) \right)_i \log \frac{\exp([\mathbf{s}(x)]_i)}{\sum_{j=1}^m \exp([\mathbf{s}(x)]_j)} \right]
\end{aligned}$$

1060 By leveraging Lemma B.1, and employing an approach analogous to that used in the proof of
1061 Proposition 3, it is clear that we can obtain

$$[\mathbf{s}^*(x)]_i = \sum_{a \in \mathcal{A}} \sum_{k=1}^N \mathbb{P}(A = a, K = k | x) \left(\left[\mathbf{M}^{\mu,\lambda}(a, k) \right]^\top \eta(x, a, k) \right)_i + \alpha.$$

1062 Hence,

$$\arg \max_{y \in [m]} [\mathbf{s}^*(x)]_y \subseteq \arg \max_{y \in [m]} \left(\sum_{a \in \mathcal{A}} \sum_{k=1}^N \mathbb{P}(A = a, K = k | x) \left[\mathbf{M}^{\mu,\lambda}(a, k) \right]^\top \eta(x, a, k) \right)_y.$$

1063 The unified classifier $h^*(x) \in \arg \min_{y \in [m]} [\mathbf{s}^*(x)]_y$ recovers that in Proposition 2. We have shown
1064 that the loss ℓ_k in (4) is also calibrated for the unified federated Bayes-optimal fair classifier. \square

1065 B.4 The Complete Formulation of Theorem 4 with Its Proof.

1066 In this subsection, we fully articulate Theorem 4 through Theorem 7 and Theorem 8, which together
1067 form an extended version of the result in Theorem 4. Before proceeding, we first clarify some
1068 notations and assumptions.

1069 With a little abuse of notation, let $f_{k,ens}^t(x) := f(x; \phi_k^t) + f(x; \theta^t)$, $f(x; \phi_k^t) = \text{softmax}(\mathbf{s}_k(x; \phi_k^t))$
1070 and $f(x; \theta^t) = \text{softmax}(\mathbf{s}_k(x; \theta^t))$. The local objective for

$$L_k^t(f(x; \theta^t)) := - \sum_{i=1}^{n_k} \sum_{y'=1}^m \bar{\mathbf{M}}_{y_i, y'}^{\lambda, \mu^t}(a_i, k) \log[f(x_i; \theta^t)], \quad k \in [N],$$

1071 which is similar to $L_k^t(f(x; \phi_k^t))$ and $L_k^t(f_{k,ens}^t(x))$.

1072 **Assumption 2.** The local loss function L_1^t, \dots, L_N^t are convex, β -smooth and bounded by B_L to
1073 model parameters ϕ_k and θ , $t \in [T]$.

Assumption 3. Let $\mathcal{B}_k^{t,r}$ be sampled from the k -th device's local data uniformly at random. The
variance of stochastic gradients in each client is bounded:

$$\mathbb{E} \left\| \nabla_\theta L_k^t(f(x; \theta); \mathcal{B}_k^{t,r}) - \nabla L_k^t(f(x; \theta)) \right\|^2 \leq \sigma^2$$

1074 for $k \in [N], t \in [T]$.

1075 Assumption 2 and 3 are standard in the convergence analysis of federated model [47, 45, 66]. Now
1076 we present Theorem 7 and Theorem 8, which together constitute an extended form of Theorem 4.

1077 **Theorem 7.** Under assumptions 2 and 3, for the ensemble personalized models, denoting $\theta^* :=$
 1078 $\arg \min_{\theta} \sum_{t=1}^T \sum_{k=1}^N p_k L_k^t(f(x; \theta))$ and $B_k = R\beta B_L + \frac{3}{2}\beta B_L + \frac{3}{4}\sigma^2$, the following cumulative
 1079 global regret upper bound of all clients is guaranteed:

$$\frac{1}{T} \sum_{t=1}^T \sum_{k=1}^N \hat{p}_k \mathbb{E}[L_k^t(f_{k,ens}^t) - L_k^t(f(x; \theta^*))] \leq \frac{\|\theta^*\|^2}{2\eta RT} + \eta B_k + \frac{\log(2)}{\eta_w T} + \eta_w B_L$$

1080 while denoting $\phi_k^* := \arg \min_{\phi_k} \sum_{t=1}^T L_k^t(f(x; \phi_k))$, the k -th client achieves the following person-
 1081 alized regret upper bound:

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}[L_k^t(f_{k,ens}^t) - L_k^t(f(x; \phi_k^*))] \leq \frac{\|\phi_k^*\|^2}{2\eta RT} + \eta B_k + \frac{\log(2)}{\eta_w T} + \eta_w B_L.$$

1082 **Theorem 8.** Suppose that personalized models achieve a ρ_t -approximate optimal response at iteration
 1083 t , namely $\hat{\mathcal{L}}(\mathbf{h}^t, \lambda^t, \mu^t) \leq \min_{\mathbf{h}} \hat{\mathcal{L}}(\mathbf{h}, \lambda^t, \mu^t) + \rho_t$, denoting $\bar{\rho} = \sum_{t=1}^T \rho_t / T$, then the sequences of
 1084 model and bounded dual parameters comprise an approximate mixed Nash equilibrium:

$$\max_{\lambda^*, \mu^*} \frac{1}{T} \sum_{t=1}^T \hat{\mathcal{L}}(\mathbf{h}^t, \lambda^*, \mu^*) - \inf_{\mathbf{h}^*} \frac{1}{T} \sum_{t=1}^T \hat{\mathcal{L}}(\mathbf{h}^*, \lambda^t, \mu^t) \leq \epsilon = \bar{\rho} + 16B_d^2 \sqrt{\frac{1}{T}}. \quad (9)$$

1085 B.4.1 Proof of Theorem 7

1086 The proof of Theorem 7 comprises proofs of the global regret bound, and the local regret bound.

1087 **(1) Global regret upper bound.** In Algorithm 1, the model parameter is updated for R iterations
 1088 locally. Therefore, for any $\theta \in \Theta$,

$$\mathbb{E} \|\theta^{t+1} - \theta\|^2 = \mathbb{E} \left\| \sum_{k=1}^N \hat{p}_k \theta_k^{t,R} - \theta \right\|^2 \leq \sum_{k=1}^N \hat{p}_k \mathbb{E} \|\theta_k^{t,R} - \theta\|^2. \quad (10)$$

1089 Denoting $g_k^{t,r} = \nabla L_k^t(f(x; \theta_k^{t,r}))$ and $G_k^{t,r} = \nabla L_k^t(f(x; \theta_k^{t,r}); \mathcal{B}_k^{t,r})$, the local update can be written
 1090 as

$$\begin{aligned} \mathbb{E} \|\theta_k^{t,r+1} - \theta\|^2 &= \mathbb{E} \|\theta_k^{t,r} - \eta G_k^{t,r} - \theta\|^2 \\ &= \mathbb{E} \|\theta_k^{t,r} - \theta\|^2 - 2\eta \mathbb{E}[\mathbb{E}[\langle G_k^{t,r}, \theta_k^{t,r} - \theta \rangle \mid \theta_k^{t,r}]] + \eta^2 \mathbb{E} \|G_k^{t,r}\|^2 \\ &\leq \mathbb{E} \|\theta_k^{t,r} - \theta\|^2 - 2\eta \mathbb{E}[\langle g_k^{t,r}, \theta_k^{t,r} - \theta \rangle] + \eta^2 (\mathbb{E} \|g_k^{t,r}\|^2 + \sigma^2). \end{aligned}$$

1091 Summarizing the inequality for $r = 0, \dots, R-1$, it shows that

$$\mathbb{E} \|\theta_k^{t,R} - \theta\|^2 = \mathbb{E} \|\theta_k^t - \theta\|^2 - 2\eta \sum_{r=0}^{R-1} \mathbb{E}[\langle g_k^{t,r}, \theta_k^{t,r} - \theta \rangle] + \eta^2 \sum_{r=0}^{R-1} (\mathbb{E} \|g_k^{t,r}\|^2 + \sigma^2). \quad (11)$$

1092 By convexity, we have

$$\begin{aligned} \sum_{r=0}^{R-1} \langle g_k^{t,r}, \theta_k^{t,r} - \theta \rangle &\geq \sum_{r=0}^{R-1} L_k^t(f(x; \theta_k^{t,r})) - L_k^t(f(x; \theta)) \\ &= \sum_{r=0}^{R-1} L_k^t(f(x; \theta_k^{t,r})) - L_k^t(f(x; \theta^t)) + L_k^t(f(x; \theta^t)) - L_k^t(f(x; \theta)) \quad (12) \end{aligned}$$

1093 By the β -smoothness, it indicates that $\|g_k^{t,r}\|^2 \leq 2\beta B_L$, and then

$$\begin{aligned} \mathbb{E}[L_k^t(f(x; \theta_k^{t,r+1}))] &\geq \mathbb{E}[L_k^t(f(x; \theta_k^{t,r}))] - \eta \mathbb{E}[\langle g_k^{t,r}, \theta_k^{t,r+1} - \theta_k^{t,r} \rangle] - \frac{\beta}{2} \|\theta_k^{t,r+1} - \theta_k^{t,r}\|^2 \\ &= \mathbb{E}[L_k^t(f(x; \theta_k^{t,r}))] - \eta \mathbb{E}[\langle g_k^{t,r}, G_k^{t,r} \rangle] - \frac{\beta \eta^2}{2} \|G_k^{t,r}\|^2 \\ &= \mathbb{E}[L_k^t(f(x; \theta_k^{t,r}))] - \eta \mathbb{E} \|g_k^{t,r}\|^2 - \frac{\beta \eta^2}{2} (\mathbb{E} \|g_k^{t,r}\|^2 + \sigma^2) \\ &\geq \mathbb{E}[L_k^t(f(x; \theta_k^{t,r}))] - \left(\eta + \frac{\beta \eta^2}{2} \right) 2\beta B_L - \frac{\beta \eta^2}{2} \sigma^2 \end{aligned}$$

1094 Summing up over $r = 0, \dots, r'$, it presents that

$$\mathbb{E}[L_k^t(f(x; \theta_k^{t,r+1}))] - L_k^t(f(x; \theta)) \geq -(2 + \eta\beta)\beta\eta B_L r' - \frac{1}{2}\beta\eta^2\sigma^2 r'$$

1095 Hence, summing up over $r' = 0, \dots, R-1$ again, we have

$$\sum_{r=0}^{R-1} \mathbb{E}[L_k^t(f(x; \theta_k^{t,r}))] - L_k^t(f(x; \theta^t)) \geq -((2 + \eta\beta)B_L - \frac{1}{2}\eta\sigma^2) \frac{\beta\eta R(R-1)}{2} \quad (13)$$

1096 Combining (11), (12) and (13), and let $\eta \leq \frac{1}{\beta R}$, we obtain

$$\begin{aligned} \mathbb{E} \left\| \theta_k^{t,R} - \theta \right\|^2 &\leq \mathbb{E} \left\| \theta^t - \theta \right\|^2 - 2\eta R [L_k^t(f(x; \theta^t)) - L_k^t(f(x; \theta))] \\ &\quad + 2\eta^2 R^2 \beta B_L + 3\eta^2 R \beta B_L + \frac{3}{2}\eta^2 R \sigma^2. \end{aligned} \quad (14)$$

1097 From (10), we know that

$$\begin{aligned} \mathbb{E} \left\| \theta^{t+1} - \theta \right\|^2 &\leq \mathbb{E} \left\| \theta^t - \theta \right\|^2 - 2\eta R \sum_{k=1}^N \hat{p}_{a,k} [L_k^t(f(x; \theta^t)) - L_k^t(f(x; \theta))] \\ &\quad + 2\eta^2 R^2 \beta B_L + 3\eta^2 R \beta B_L + \frac{3}{2}\eta^2 R \sigma^2 \end{aligned}$$

1098 Summing over time and dividing both sides by $\frac{1}{2\eta RT}$, we obtain

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \sum_{k=1}^N \hat{p}_{a,k} \mathbb{E}[L_k^t(f(x; \theta^t)) - L_k^t(f(x; \theta^*))] \\ \leq \frac{\mathbb{E} \left\| \theta^0 - \theta \right\|^2 - \mathbb{E} \left\| \theta^{T+1} - \theta \right\|^2}{2\eta RT} + \eta(R\beta B_L + \frac{3}{2}\beta B_L + \frac{3}{4}\sigma^2). \end{aligned}$$

1099 Plugging in $\theta = \theta^*$ and $\theta^0 = 0$ and considering the fact that $\theta^{T+1} - \theta \geq 0$, the result turns to

$$\frac{1}{T} \sum_{t=1}^T \sum_{k=1}^N \hat{p}_{a,k} \mathbb{E}[L_k^t(f(x; \theta^t)) - L_k^t(f(x; \theta^*))] \leq \frac{\|\theta^*\|^2}{2\eta RT} + \eta(R\beta B_L + \frac{3}{2}\beta B_L + \frac{3}{4}\sigma^2). \quad (15)$$

1100 Consider the update rule of ensemble weight w_k^t in Algorithm 1,

$$w_k^{t+1} = \frac{1}{1 + \mathcal{W}_k^t(w_k^t)} = \frac{w_k^t \exp(-\eta_w L_k^t(\theta_k))}{w_k^t \exp(-\eta_w L_k^t(\theta^t)) + (1 - w_k^t) \exp(-\eta_w L_k^t(\phi_k^t))}.$$

1101 Here, the update can be viewed as exponentiated gradient descent on the normalized weight vector
1102 $\mathbf{w}_k^t = (w_{k,1}^t, w_{k,2}^t) \in \Delta_2$, and $w_{k,i}^t \mathbb{P}^{opt} \exp(-\eta_w z_{t,i}^k)$, $i = 1, 2$, where $z_{t,1}^k = L_k^t(f(x; \theta^t))$, $z_{t,2}^k =$
1103 $L_k^t(f(x; \phi_k^t))$. A well-known regret bound in online learning [63, 64] shows that, for any $\mathbf{u} =$
1104 $(u_1, u_2) \in \Delta_2$,

$$\begin{aligned} \sum_{t=1}^T w_k^t \mathbb{E}[L_k^t(f(x; \theta^t))] + (1 - w_k^t) \mathbb{E}[L_k^t(f(x; \phi_k^t))] - \sum_{t=1}^T (u_1 \mathbb{E}[L_k^t(f(x; \theta^t))] + u_2 \mathbb{E}[L_k^t(f(x; \phi_k^t))]) \\ \leq \frac{\log(2)}{\eta_w} + \eta_w T B_L. \end{aligned} \quad (16)$$

1105 By the convexity of L_k^t , we have $\sum_{t=1}^T L_k^t(f_{k,ens}^t) \leq \sum_{t=1}^T w_k^t L_k^t(f(x; \theta^t)) + (1 - w_k^t) L_k^t(f(x; \phi_k^t))$.

1106 Plugging in $u_1 = 1, u_2 = 0$, it presents that

$$\sum_{t=1}^T \mathbb{E}[L_k^t(f_{k,ens}^t) - L_k^t(f(x; \theta^t))] \leq \frac{\log(2)}{\eta_w} + \eta_w T B_L. \quad (17)$$

1107 Weighted summing (17) over all clients and dividing both sides by T , we obtain

$$\frac{1}{T} \sum_{t=1}^T \sum_{k=1}^N \hat{p}_k \mathbb{E}[L_k^t(f_{k,ens}^t) - L_k^t(f(x; \theta^t))] \leq \frac{\log(2)}{\eta_w T} + \eta_w B_L. \quad (18)$$

1108 Combining (18) and (15), and denoting $B_k = R\beta B_L + \frac{3}{2}\beta B_L + \frac{3}{4}\sigma^2$, we obtain

$$\frac{1}{T} \sum_{t=1}^T \sum_{k=1}^N \hat{p}_k \mathbb{E}[L_k^t(f_{k,ens}^t) - L_k^t(f(x; \theta^*))] \leq \frac{\|\theta^*\|^2}{2\eta RT} + \eta B_k + \frac{\log(2)}{\eta_w T} + \eta_w B_L. \quad (19)$$

1109 Thus, we finish the proof of the global regret upper bound.

1110 **(2) Local regret upper bound.** Plugging in $u_1 = 0, u_2 = 1$ in (16), it presents that

$$\sum_{t=1}^T \mathbb{E}[L_k^t(f_{k,ens}^t) - L_k^t(f(x; \phi_k^t))] \leq \frac{\log(2)}{\eta_w} + \eta_w T B_L \quad (20)$$

1111 Following the proof technique of global regret upper bound, from (14), since $\phi_k^{t,R} = \phi_k^{t+1}$, and
 1112 making $\eta \leq \frac{1}{\beta R}$, we have for any ϕ_k ,

$$\begin{aligned} \mathbb{E} \|\phi_k^{t+1} - \phi_k\|^2 &\leq \mathbb{E} \|\phi_k^t - \phi_k\|^2 - 2\eta R [L_k^t(f(x; \phi_k^t)) - L_k^t(f(x; \phi_k))] \\ &\quad + 2\eta^2 R^2 \beta B_L + 3\eta^2 R \beta B_L + \frac{3}{2}\eta^2 R \sigma^2. \end{aligned} \quad (21)$$

1113 Combining (20) and (21), and plugging in $\phi_k = \phi_k^*$ and $\phi_k^0 = 0$ denoting $B_k = R\beta B_L + \frac{3}{2}\beta B_L + \frac{3}{4}\sigma^2$,
 1114 the result turns to

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}[L_k^t(f_{k,ens}^t) - L_k^t(f(x; \phi_k^*))] \leq \frac{\|\phi_k^*\|^2}{2\eta RT} + \eta B_k + \frac{\log(2)}{\eta_w T} + \eta_w B_L. \quad (22)$$

1115 Thus, we finish the proof of the local regret upper bound. \square

1116 B.4.2 Proof of Theorem 8

1117 The proof of Theorem 8 relies on Lemma B.2.

Lemma B.2. [63] Let $f^1, f^2, \dots : \Lambda \rightarrow \mathbb{R}$ be a sequence of convex functions that we wish to minimize on a compact convex set Λ . Define the bound of the convex set $B_d \geq \max_{\lambda \in \Lambda} \|\lambda\|_2$, and $B_G \geq \|\nabla f^t(\lambda^t)\|_2$ is a uniform upper bound on the norms of the subgradients. Suppose that we perform T iterations of the following update, starting from $\lambda^{(1)} = \arg\min_{\lambda \in \Lambda} \|\lambda\|_1$:

$$\Lambda^t = \Pi_\Lambda \left(\lambda^{(t)} - \eta \nabla f^t(\lambda^{(t)}) \right)$$

where $\nabla f^t(\lambda^{(t)}) \in \partial f^t(\lambda^{(t)})$ is a subgradient of f^t at (λ) , and Π_Λ projects its argument onto Λ w.r.t. the Euclidean norm. Then:

$$\frac{1}{T} \sum_{t=1}^T f^t(\lambda^{(t)}) - \frac{1}{T} \sum_{t=1}^T f^t(\lambda^*) \leq \frac{B_d^2}{2\eta} + \eta T B_G^2$$

1118 where $\lambda^* \in \Lambda$ is an arbitrary reference vector.

1119 *Proof of Theorem 8.* Consider the empirical form of the Lagrangian function $\hat{\mathcal{L}}(\mathbf{h}, \lambda, \mu)$,

$$\begin{aligned} \hat{\mathcal{L}}(\mathbf{h}, \lambda, \mu) &= \hat{\mathcal{R}}(\mathbf{h}) + (\lambda^{(1)} - \lambda^{(2)})^\top (\hat{\mathcal{G}}^g(\mathbf{h}) - \xi^g) + \sum_{k=1}^N (\mu^{1,k} - \mu^{2,k})^\top (\hat{\mathcal{G}}^{l,k}(\mathbf{h}) - \xi^{l,k}), \\ &= \sum_{k=1}^N \hat{p}_k \frac{1}{n_k} \sum_{i=1}^{n_k} e_{y_{k,i}}^\top [\mathbf{1} - \hat{\mathbf{M}}^{\lambda, \mu}(a_{k,i}, k)] h_k(x_{k,i}). \end{aligned}$$

1120 where $\widehat{\mathbf{M}}^{\lambda, \mu}(a, k) := \mathbf{I} - \frac{1}{\widehat{\rho}_{a, k}} \left[\sum_{u_g \in \mathcal{U}_g} (\lambda_{u_g}^{(1)} - \lambda_{u_g}^{(2)}) \widehat{\mathbf{D}}_{u_g}^{a, k} - \sum_{u_{l, k} \in \mathcal{U}_{l, k}} (\mu_{k, u_{l, k}}^{(1)} - \mu_{k, u_{l, k}}^{(2)}) \widehat{\mathbf{D}}_{u_{l, k}}^{a, k} \right]$.

1121 It is clear that the inner problem is linear to classifiers in the empirical case.

1122 From the definition in Section 4, we have $\|\lambda\|_1 \leq B_d, \|\mu\|_1 \leq B_d$. Since the norm of fairness metrics
1123 is less than 2, setting the step size $\eta_d = B_d/\sqrt{T}$, by Lemma B.2,

$$\max_{\lambda^*, \mu^*} \frac{1}{T} \sum_{t=1}^T \widehat{\mathcal{L}}(\mathbf{h}^t, \lambda^*, \mu^*) - \frac{1}{T} \sum_{t=1}^T \widehat{\mathcal{L}}(\mathbf{h}^t, \lambda^t, \mu^t) \leq \frac{2B_d^2}{\eta_d} + 4\eta_d T = 16B_d^2 \sqrt{\frac{1}{T}}, \quad (23)$$

1124 where λ^*, μ^* are the optimal dual parameters satisfying $\|\lambda\|_1 \leq B_d, \|\mu\|_1 \leq B_d$.

1125 On the other hand, according to the sub-optimal assumption on the classifier \mathbf{h} , we have

$$\frac{1}{T} \sum_{t=1}^T \widehat{\mathcal{L}}(\mathbf{h}^t, \lambda^t, \mu^t) - \inf_{\mathbf{h}^*} \frac{1}{T} \sum_{t=1}^T \widehat{\mathcal{L}}(\mathbf{h}^*, \lambda^t, \mu^t) \leq \frac{1}{T} \sum_{t=1}^T \widehat{\mathcal{L}}(\mathbf{h}^t, \lambda^t, \mu^t) - \frac{1}{T} \sum_{t=1}^T \inf_{\mathbf{h}^*} \widehat{\mathcal{L}}(\mathbf{h}^*, \lambda^t, \mu^t) \leq \bar{\rho}, \quad (24)$$

1126 where $\bar{\rho} := \sum_{t=1}^T \rho_t / T$. Combining (23) and (24), the result shows that

$$\max_{\lambda^*, \mu^*} \frac{1}{T} \sum_{t=1}^T \widehat{\mathcal{L}}(\mathbf{h}^t, \lambda^*, \mu^*) - \inf_{\mathbf{h}^*} \frac{1}{T} \sum_{t=1}^T \widehat{\mathcal{L}}(\mathbf{h}^*, \lambda^t, \mu^t) \leq \bar{\rho} + 16B_d^2 \sqrt{\frac{1}{T}}. \quad (25)$$

1127 Let $\bar{\mathbf{h}} := \frac{1}{T} \sum_{t=1}^T \mathbf{h}^t$ with $\bar{h}_k := \frac{1}{T} \sum_{t=1}^T h_k^t, k \in [N]$, and let $\bar{\lambda} := \frac{1}{T} \sum_{t=1}^T \lambda^t, \bar{\mu} := \frac{1}{T} \sum_{t=1}^T \mu^t$
1128 denote the point-wise average of dual parameters. Therefore, due to the linearity of empirical
1129 Lagrange function to classifiers and dual parameters, (25) can be formulated as

$$\max_{\lambda^*, \mu^*} \widehat{\mathcal{L}}(\bar{\mathbf{h}}, \lambda^*, \mu^*) - \inf_{\mathbf{h}^*} \widehat{\mathcal{L}}(\mathbf{h}^*, \bar{\lambda}, \bar{\mu}) \leq \bar{\rho} + 16B_d^2 \sqrt{\frac{1}{T}}, \quad (26)$$

1130 which presents the approximate mixed Nash equilibrium of the stochastic saddle-point problem. \square

1131 B.5 Generalization Error For In-processing Algorithm

1132 We begin by introducing some notations and simplifications, which are commonly employed in
1133 generalization analyses of FL [35, 61]. Without loss of generalization, let $n = n_1 = \dots =$
1134 n_k present the sample number in local datasets. For any class $\mathcal{H} = \{h : \mathcal{X} \rightarrow [m]\}$, denote
1135 $\mathcal{H}_y = \{\mathbb{I}\{h(x) = y\} : h \in \mathcal{H}\}$ and the maximal Vapnik-Chervonenkis dimension [62], $VC(\mathcal{H}) :=$
1136 $\max_{y \in [m]} VC(\mathcal{H}_y)$.

1137 **Theorem 9.** *If classifiers $\bar{\mathbf{h}} = (\bar{h}_1, \dots, \bar{h}_k)$ with dual parameters $(\bar{\lambda}, \bar{\mu})$ form a ϵ -saddle point
1138 of empirical Lagrangian $\widehat{\mathcal{L}}(\mathbf{h}, \lambda, \mu)$, and an optimal solution $\mathbf{h}^* \in \mathcal{H}$ satisfies both global and
1139 local fairness constraints, denoting $\nu(n, \mathcal{H}, \delta) = 2\sqrt{\frac{2VC(\mathcal{H}) \log(n+1)}{n}} + \sqrt{\frac{2 \log(m^2 N / \delta)}{n}}$, $B_g =$
1140 $\max_{a \in \mathcal{A}, k \in [N]} \|\mathbf{D}_{u_g}^{a, k}\|_1, \Omega_n^g = \max_{a \in \mathcal{A}, k \in [N]} \|\mathbf{D}_{u_g}^{a, k} - \widehat{\mathbf{D}}_{u_g}^{a, k}\|_\infty, \Omega_n^p := \sum_{k=1}^N |p_k - \widehat{p}_k|$, and $B_{l, k} =$
1141 $\max_{a \in \mathcal{A}} \|\mathbf{D}_{u_{l, k}}^{a, k}\|_1, \Omega_n^{l, k} = \max_{a \in \mathcal{A}} \|\mathbf{D}_{u_{l, k}}^{a, k} - \widehat{\mathbf{D}}_{u_{l, k}}^{a, k}\|_\infty, k \in [N]$, then with probability at least $1 - \delta$,*

$$\begin{aligned} |\mathcal{D}^g(\bar{\mathbf{h}})| &\leq \xi^g + \nu(n, \mathcal{H}, \delta/|\mathcal{A}||\mathcal{U}_g|)|\mathcal{A}|NB_g + \Omega_n^g + \frac{1+2\epsilon}{B_d}, \\ |\mathcal{D}^{l, k}(\bar{\mathbf{h}})| &\leq \xi^{l, k} + \nu(n, \mathcal{H}, N\delta/|\mathcal{A}||\mathcal{U}_{l, k}|)|\mathcal{A}|B_{l, k} + \Omega_n^{l, k} + \frac{1+2\epsilon}{B_d} \\ \mathcal{R}(\bar{\mathbf{h}}) &\leq \mathcal{R}(\mathbf{h}^*) + 2m\Omega_n^p + 2m\nu(n, \mathcal{H}, \delta/2) + 2\epsilon. \end{aligned}$$

1142 The proof of Theorem 9 relies on the following lemma.

1143 **Lemma B.3.** *Let $\mathcal{H} : \mathcal{X} \rightarrow [m]$, \mathcal{D} a distribution over $\mathcal{X} \times \Delta_m$, of which $\{x_i, y_i\}_{i=1}^n$ are i.i.d samples.
1144 Denoting $\mathcal{H}_y = \{\mathbb{I}\{h(x) = y\} : h \in \mathcal{H}\}$ and $VC(\mathcal{H}) = \max_{y \in [m]} VC(\mathcal{H}_y)$, then with probability
1145 at least $1 - \delta$, for $\forall i, j \in [m]$,*

$$\sup_{h \in \mathcal{F}_{\mathcal{H}}} |\mathbf{C}_{i, j}(h) - \widehat{\mathbf{C}}_{i, j}(h)| \leq 2\sqrt{\frac{2VC(\mathcal{H}) \log(n+1)}{n}} + \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}.$$

1146 where $\mathcal{F}_{\mathcal{H}} := \{f(x) = \sum_{j=1}^N \alpha_j h_j(x) : \alpha \in \Delta_N, h_j \in \mathcal{H}, j \in [m]\}$.

1147 *Proof of Lemma B.3.* Let $\ell_{i,j}(x, y; h) = \mathbb{I}(y = i \wedge h(x) = j)$. Then we have $\mathbf{C}_{i,j}(h) =$
 1148 $\mathbb{E}[\ell_{i,j}(x, y; h)]$ and $\widehat{\mathbf{C}}_{i,j}(h) = \frac{1}{n} \sum_{i=1}^n \ell_{i,j}(x_i, y_i; h)$. Hence, according to the classical result with
 1149 respect to cost sensitive binary classification [7], with probability at least $1 - \delta$,

$$\sup_{h \in \mathcal{F}_{\mathcal{H}}} |\mathbf{C}_{i,j}(h) - \widehat{\mathbf{C}}_{i,j}(h)| \leq 2\sqrt{\frac{2VC(\mathcal{H}_j) \log(n+1)}{n}} + \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}.$$

1150 By the definition of $VC(\mathcal{H})$, it achieves the generalization bound.

1151 B.5.1 Proof of Theorem 9

1152 Let the optimal solution \mathbf{h}^* minimize the risk $\mathcal{R}(\mathbf{h})$ subjected to global and local fairness constraints
 1153 $|\mathcal{D}^g(\mathbf{h}^*)| \leq \xi^g$, $|\mathcal{D}^{k,l}(\mathbf{h}^*)| \leq \xi^{k,l}$. With the properties of the saddle point, it is clear that

$$\widehat{\mathcal{L}}(\bar{\mathbf{h}}, \bar{\lambda}, \bar{\mu}) \leq \widehat{\mathcal{L}}(\mathbf{h}, \bar{\lambda}, \bar{\mu}) + \epsilon, \quad \forall \mathbf{h} \in \mathcal{H}, \quad (27)$$

$$\widehat{\mathcal{L}}(\bar{\mathbf{h}}, \bar{\lambda}, \bar{\mu}) \geq \widehat{\mathcal{L}}(\bar{\mathbf{h}}, \lambda, \mu) - \epsilon, \quad \forall \|\lambda\|_1, \|\mu\|_1 \leq B_d. \quad (28)$$

1154 Considering the global fairness constraints, we first explore its concentration, for any $h \in \mathcal{H}$,

$$\begin{aligned} \mathcal{D}_{u_g}^g(\mathbf{h}) - \widehat{\mathcal{D}}_{u_g}^g(\mathbf{h}) &= \sum_{k=1}^N \sum_{a \in \mathcal{A}} \langle \mathbf{D}_{u_g}^{a,k}, \mathbf{C}^{a,k}(h_k) \rangle - \langle \widehat{\mathbf{D}}_{u_g}^{a,k}, \widehat{\mathbf{C}}^{a,k}(h_k) \rangle \\ &= \sum_{k=1}^N \sum_{a \in \mathcal{A}} \langle \mathbf{D}_{u_g}^{a,k}, \mathbf{C}^{a,k}(h_k) - \widehat{\mathbf{C}}^{a,k}(h_k) \rangle + \langle \mathbf{D}_{u_g}^{a,k} - \widehat{\mathbf{D}}_{u_g}^{a,k}, \widehat{\mathbf{C}}^{a,k}(h_k) \rangle \\ &\leq \sum_{k=1}^N \sum_{a \in \mathcal{A}} \|\mathbf{D}_{u_g}^{a,k}\|_1 \|\mathbf{C}^{a,k}(h_k) - \widehat{\mathbf{C}}^{a,k}(h_k)\|_{\infty} + \|\mathbf{D}_{u_g}^{a,k} - \widehat{\mathbf{D}}_{u_g}^{a,k}\|_{\infty} \|\widehat{\mathbf{C}}^{a,k}(h_k)\|_1. \end{aligned}$$

1155 The last inequality is by the Holder's inequality. Let $\ell_{i,j}^{a',k}(x, y, a; h) = \mathbb{I}(y = i \wedge h(x) = j \wedge a = a')$.
 1156 Then we have $\mathbf{C}_{i,j}^{a',k}(h) = \mathbb{E}[\ell_{i,j}^{a',k}(x, y, a; h)]$ and $\widehat{\mathbf{C}}_{i,j}^{a',k}(h) = \frac{1}{n} \sum_{i=1}^n \ell_{i,j}^{a',k}(x_i, y_i; h)$. By taking a
 1157 union bound in Lemma B.3, we have that with probability at least $1 - \delta$,

$$\sup_{h \in \mathcal{F}_{\mathcal{H}}} \max_{k \in [N]} \max_{a \in \mathcal{A}} \|\mathbf{C}^{a,k}(h) - \widehat{\mathbf{C}}^{a,k}(h)\|_{\infty} \leq 2\sqrt{\frac{2VC(\mathcal{H}) \log(n+1)}{n}} + \sqrt{\frac{2 \log(m^2 |\mathcal{A}| N / \delta)}{n}}.$$

1158 Since $\|\widehat{\mathbf{C}}^{a,k}(h_k)\|_1 = 1$, taking union bound again, denoting $\nu(n, \mathcal{H}, \delta) = 2\sqrt{\frac{2VC(\mathcal{H}) \log(n+1)}{n}} +$
 1159 $\sqrt{\frac{2 \log(m^2 N / \delta)}{n}}$, it turns out that with probability at least $1 - \delta$,

$$\mathcal{D}_{u_g}^g(\mathbf{h}) - \widehat{\mathcal{D}}_{u_g}^g(\mathbf{h}) \leq \sum_{k=1}^N \sum_{a \in \mathcal{A}} \nu(n, \mathcal{H}, \delta / |\mathcal{A}|) \|\mathbf{D}_{u_g}^{a,k}\|_1 + \|\mathbf{D}_{u_g}^{a,k} - \widehat{\mathbf{D}}_{u_g}^{a,k}\|_{\infty} \quad (29)$$

$$\leq \nu(n, \mathcal{H}, \delta / |\mathcal{A}|) |\mathcal{A}| N B_g + \Omega_n^g. \quad (30)$$

1160 Next, we consider the optimality. Denoting $u_g^* := \arg \max_{u_g \in \mathcal{U}_g} |\widehat{\mathcal{D}}_{u_g}^g(\bar{\mathbf{h}})|$, then we have

$$B_d(\widehat{\mathcal{D}}_{u_g^*}^g(\bar{\mathbf{h}}) - \xi^g) = \widehat{\mathcal{L}}(\bar{\mathbf{h}}, B e_{u_g^*}^1, 0) - \widehat{R}(\bar{\mathbf{h}}) \leq \widehat{\mathcal{L}}(\bar{\mathbf{h}}, \bar{\lambda}, \bar{\mu}) - \widehat{R}(\bar{\mathbf{h}}) + \epsilon, \quad (31)$$

1161 where $e_{u_g^*}^1$ defines as the basis vector with 1 at the position of $\lambda_{u_g^*}^1$. Let \mathbf{h} satisfy the fairness
 1162 constraints. With (27), we obtain

$$\widehat{\mathcal{L}}(\bar{\mathbf{h}}, \bar{\lambda}, \bar{\mu}) - \widehat{R}(\bar{\mathbf{h}}) \leq \widehat{\mathcal{L}}(\mathbf{h}, \bar{\lambda}, \bar{\mu}) - \widehat{R}(\bar{\mathbf{h}}) + \epsilon \leq \widehat{R}(\mathbf{h}) - \widehat{R}(\bar{\mathbf{h}}) + \epsilon. \quad (32)$$

1163 Combining (31) and (32), it shows that

$$\widehat{\mathcal{D}}_{u_g^*}^g(\bar{\mathbf{h}}) - \xi^g \leq \frac{\widehat{R}(\mathbf{h}) - \widehat{R}(\bar{\mathbf{h}}) + 2\epsilon}{B_d} \leq \frac{1 + 2\epsilon}{B_d}. \quad (33)$$

1164 Therefore, the result shows that $\max_{u_g \in \mathcal{U}_g} |\hat{\mathcal{D}}_{u_g}^g(\bar{\mathbf{h}}) - \xi^g| \leq \frac{1+2\epsilon}{B_d}$.

1165 Now we consider the generalization error for the empirically optimal classifier $\bar{\mathbf{h}}$, with probability at
1166 least $1 - \delta$,

$$|\mathcal{D}_{u_g}^g(\bar{\mathbf{h}}) - \xi^g| \leq |\mathcal{D}_{u_g}^g(\bar{\mathbf{h}}) - \hat{\mathcal{D}}_{u_g}^g(\bar{\mathbf{h}})| + |\hat{\mathcal{D}}_{u_g}^g(\bar{\mathbf{h}}) - \xi^g| \quad (34)$$

$$\leq \nu(n, \mathcal{H}, \delta/|\mathcal{A}|) |\mathcal{A}| NB_g + \Omega_n^g + \frac{1+2\epsilon}{B_d}. \quad (35)$$

1167 Taking the union bound over $u_g \in \mathcal{U}_g$, we have that with probability at least $1 - \delta$,

$$|\mathcal{D}^g(\bar{\mathbf{h}}) - \xi^g| \leq \nu(n, \mathcal{H}, \delta/|\mathcal{A}||\mathcal{U}_g|) |\mathcal{A}| NB_g + \Omega_n^g + \frac{1+2\epsilon}{B_d}. \quad (36)$$

1168 For local fairness constraints, $|\mathcal{D}_{l,k}| \leq \xi^{l,k}$, following the similar proof procedures as local fairness
1169 constraints, we have that

$$|\mathcal{D}^{l,k}(\bar{\mathbf{h}}) - \xi^{l,k}| \leq \nu(n, \mathcal{H}, N\delta/|\mathcal{A}||\mathcal{U}_{l,k}|) |\mathcal{A}| B_{l,k} + \Omega_n^{l,k} + \frac{1+2\epsilon}{B_d}. \quad (37)$$

1170 For risk metric $\mathcal{R}(\mathbf{h})$, it presents that

$$\mathcal{R}(\bar{\mathbf{h}}) - \mathcal{R}(\mathbf{h}^*) = \mathcal{R}(\bar{\mathbf{h}}) - \hat{\mathcal{R}}(\bar{\mathbf{h}}) + \hat{\mathcal{R}}(\bar{\mathbf{h}}) - \hat{\mathcal{R}}(\mathbf{h}^*) + \hat{\mathcal{R}}(\mathbf{h}^*) - \mathcal{R}(\mathbf{h}^*). \quad (38)$$

1171 By (27) and (28),

$$\hat{\mathcal{R}}(\bar{\mathbf{h}}) - \hat{\mathcal{R}}(\mathbf{h}^*) \leq \hat{\mathcal{L}}(\bar{\mathbf{h}}, \mathbf{0}, \mathbf{0}) - \hat{\mathcal{L}}(\mathbf{h}^*, \bar{\lambda}, \bar{\mu}) \leq \hat{\mathcal{L}}(\bar{\mathbf{h}}, \bar{\lambda}, \bar{\mu}) + \epsilon - \hat{\mathcal{L}}(\bar{\mathbf{h}}, \bar{\lambda}, \bar{\mu}) + \epsilon = 2\epsilon. \quad (39)$$

1172 Since we have $\hat{\mathcal{R}}(\mathbf{h}) = 1 - \sum_{k=1}^N \hat{p}_k \langle \mathbf{I}, \mathbf{C}^k(h_k) \rangle$, it presents that

$$\begin{aligned} \hat{\mathcal{R}}(h) - \mathcal{R}(h) &= \sum_{k=1}^N \langle \mathbf{I}, p_k \mathbf{C}^k(h_k) - \hat{p}_k \hat{\mathbf{C}}^k(h_k) \rangle \\ &= \sum_{k=1}^N (p_k - \hat{p}_k) \langle \mathbf{I}, \mathbf{C}^k(h_k) \rangle + \sum_{k=1}^N p_k \langle \mathbf{I}, \mathbf{C}^k(h_k) - \hat{\mathbf{C}}^k(h_k) \rangle \\ &\leq m \sum_{k=1}^N |p_k - \hat{p}_k| + \sum_{k=1}^N p_k m \|\mathbf{C}^k(h_k) - \hat{\mathbf{C}}^k(h_k)\|_\infty \end{aligned}$$

1173 By taking a union bound in Lemma B.3, we have that with probability at least $1 - \delta$,

$$\sup_{h \in \mathcal{F}_{\mathcal{H}}} \max_{k \in [N]} \|\mathbf{C}^k(h) - \hat{\mathbf{C}}^k(h)\|_\infty \leq 2\sqrt{\frac{2VC(\mathcal{H}) \log(n+1)}{n}} + \sqrt{\frac{2 \log(m^2 N/\delta)}{n}}.$$

1174 Hence, denoting $\Omega_n^p := \sum_{k=1}^N |p_k - \hat{p}_k|$, we arrive that, for any $h \in \mathcal{F}_{\mathcal{H}}$,

$$\begin{aligned} \hat{\mathcal{R}}(h) - \mathcal{R}(h) &\leq N \max_{k \in [N]} |p_k - \hat{p}_k| + \sum_{k=1}^N p_k m \|\mathbf{C}^k(h_k) - \hat{\mathbf{C}}^k(h_k)\|_\infty \\ &\leq m\Omega_n^p + m\nu(n, \mathcal{H}, \delta). \end{aligned} \quad (40)$$

1175 Therefore, combining (38), (39) and (40), we obtain

$$\mathcal{R}(\bar{\mathbf{h}}) - \mathcal{R}(\mathbf{h}^*) \leq 2m\Omega_n^p + 2m\nu(n, \mathcal{H}, \delta/2) + 2\epsilon. \quad (41)$$

1176 This completes the proof. \square

1177 B.6 Proof of Theorem 5

1178 We begin by introducing some definitions and lemmas, which are useful in the proof of Theorem 5.

1179 **Definition 3.** Let V be a real vector space and let $A, B \subseteq V$. The sum of A and B is defined by

$$A + B := \{a + b \mid a \in A, b \in B\}.$$

1180 **Lemma B.4.** [5] The subdifferential of the function $F(x) = \mathbb{E}\{f(x, \omega)\}$ at a point x is given by

$$\partial F(x) = \mathbb{E}\{\partial f(x, \omega)\}$$

1181 where $f(\cdot, \omega)$ is a real-value convex function and the set $\mathbb{E}\{\partial f(x, \omega)\}$ is defined as

$$\begin{aligned} \mathbb{E}\{\partial f(x, \omega)\} &:= \int_{\Omega} \partial f(x, \omega) d\mathbb{P}(\omega) \\ &= \left\{ x^* \in \mathbb{R}^n \mid x^* = \int_{\Omega} x^*(\omega) d\mathbb{P}(\omega), x^*(\cdot) : \text{measurable}, x^*(\omega) \in \partial f(x, \omega) \text{ a.e.} \right\}. \end{aligned}$$

1182 **Lemma B.5.** [59] Let $f_1, \dots, f_m : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ be convex functions. Define $f(x) =$
 1183 $\max\{f_1(x), \dots, f_m(x)\}, \quad \forall x \in \mathbb{R}^n$. For $x_0 \in \bigcap_{i=1}^m \text{dom} f_i$, define $I(x_0) = \{i \mid f_i(x_0) = f(x_0)\}$.
 1184 Then $\partial f(x_0) = \text{conv} \bigcup_{i \in I(x_0)} \partial f_i(x_0)$.

1185 **Lemma B.6.** [59] Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a convex continuous function. We consider the minimizer x^*
 1186 of the function f over the set B . Then for x^* to be locally optimal it is necessary that

$$\partial f(x^*) + \mathcal{N}_B(x^*) \ni 0,$$

1187 where \mathcal{N}_B denotes the normal cone of set B . If $B = \mathbb{R}^d$, let $\mathcal{K} := \{k \in [d], x_k^* \neq 0\}$. Then there
 1188 exists a subgradient $\xi \in \partial f(x^*)$, such that for all $k \in [d]$ we have $\xi_k \geq 0$ and $\forall k \in \mathcal{K}, \xi_k = 0$.

1189 *Proof of Theorem 5.* From the above analysis, it follows that the Lagrange function can be written as

$$\begin{aligned} \mathcal{L}(\mathbf{h}, \lambda, \mu) &= 1 - \sum_{k=1}^N \sum_{a \in \mathcal{A}} p_{a,k} \langle \mathbf{M}^{\lambda, \mu}(a, k), \mathbf{C}^{a,k}(h_k) \rangle - \sum_{u_g \in \mathcal{U}_g} (\lambda_{u_g}^{(1)} + \lambda_{u_g}^{(2)}) \xi^g \\ &\quad - \sum_{k \in [N]} \sum_{u_{l,k} \in \mathcal{U}_{l,k}} (\mu_{k,u_{l,k}}^{(1)} + \mu_{k,u_{l,k}}^{(2)}) \xi^{l,k}. \end{aligned}$$

1190 We first consider the inner optimization problem $\min_{\mathbf{h} \in \mathcal{H}^N} \mathcal{L}(\mathbf{h}, \lambda, \mu)$, which is equivalent to opti-
 1191 mize

$$\max_{\mathbf{h} \in \mathcal{H}^N} V(\mathbf{h}, \lambda, \mu) = \sum_{k=1}^N \sum_{a \in \mathcal{A}} p_{a,k} \langle \mathbf{M}^{\lambda, \mu}(a, k), \mathbf{C}^{a,k}(h_k) \rangle.$$

1192 where $\mathbf{M}^{\lambda, \mu}(a, k) := \mathbf{I} - \frac{1}{p_{a,k}} \left[\sum_{u_g \in \mathcal{U}_g} (\lambda_{u_g}^{(1)} - \lambda_{u_g}^{(2)}) \mathbf{D}_{u_g}^{a,k} - \sum_{u_{l,k} \in \mathcal{U}_{l,k}} (\mu_{k,u_{l,k}}^{(1)} - \mu_{k,u_{l,k}}^{(2)}) \mathbf{D}_{u_{l,k}}^{a,k} \right]$.
 1193 Considering the personalized attribute-aware classifier $h_k(x, a), k \in [N]$ in post-processing, the inner
 1194 function turns to

$$\begin{aligned} V(\mathbf{h}, \lambda, \mu) &= \sum_{k=1}^N \sum_{a \in \mathcal{A}} p_{a,k} \langle \mathbf{M}^{\lambda, \mu}(a, k), \mathbf{C}^{a,k}(h_k) \rangle \\ &= \sum_{k=1}^N \sum_{a \in \mathcal{A}} p_{a,k} \int_{\mathcal{X}} [\eta(x, a, k)]^\top \mathbf{M}^{\lambda, \mu}(a, k) h_k(x, a) d\mathcal{P}_{a,k}^X \end{aligned}$$

An explicit optimal solution of personalized classifier is that

$$h_k^{\lambda, \mu}(x, a) := \arg \max_{y \in [m]} \left([\mathbf{M}^{\lambda, \mu}(a, k)]^\top \eta(x, a, k) \right)_j.$$

1195 If the maximum entry of the output vector occurs at multiple indices, one of them is randomly selected
 1196 as the predicted class. Thus, the dual problem can be formulated as

$$\begin{aligned} \min_{\lambda, \mu} H(\lambda, \mu) &:= \sum_{k \in [N]} \sum_{a \in \mathcal{A}} p_{a,k} \mathbb{E}_{X \sim \mathcal{P}_{a,k}^X} \left[\max_{y \in [m]} \left([\mathbf{M}^{\lambda, \mu}(a, k)]^\top \eta(X, a, k) \right)_y \right] + \xi^g \sum_{u_g \in \mathcal{U}_g} (\lambda_{u_g}^{(1)} + \lambda_{u_g}^{(2)}) \\ &\quad + \sum_{k \in [N]} \xi^{l,k} \sum_{u_{l,k} \in \mathcal{U}_{l,k}} (\mu_{k,u_{l,k}}^{(1)} + \mu_{k,u_{l,k}}^{(2)}). \end{aligned} \tag{42}$$

Before exploring the optimal solution of outer optimization, we first prove that the optimal dual parameter $\lambda^* \in \mathbb{R}_{\geq 0}^{2|\mathcal{U}_g|}$, $\mu^* \in \mathbb{R}_{\geq 0}^{2\sum_{k=1}^N |\mathcal{U}_{l,k}|}$ is bounded. Define the Hilbert space on $\mathcal{F} := \{f : \mathcal{X} \rightarrow \mathbb{R}^m\}$ with inner product $\langle f, g \rangle = \int_{\mathcal{X}} f^\top g d\mathcal{P}(x)$. Then the classifier space $\mathcal{H} : \mathcal{X} \rightarrow \Delta_m$ is a convex subset of \mathcal{F} . Therefore, we can also consider the topology structure on \mathcal{H} or $\mathcal{H}^{|\mathcal{A}|}$. Since we assume that $\forall \xi^g, \xi^{l,k} > 0$, the feasible set of the primal problem is non-empty, it indicates that the feasible set of the primal problem has non-empty interior for any positive $\xi^g, \xi^{l,k}$. It is clear that for $\forall \xi^g, \xi^{l,k} > 0$, the dual problem

$$\min_{\mathbf{h}} \mathcal{L}(\mathbf{h}, \lambda, \mu) = 1 - H(\lambda, \mu) \leq \mathcal{R}(\mathbf{h}) \leq \mathcal{R}_{\max}(\mathbf{h}^{fair})$$

where \mathbf{h}^{fair} denotes a classifier that satisfies fairness constraints for given $\xi^g, \xi^{l,k} > 0$. Hence, we arrive at

$$H(\lambda, \mu) \geq 1 - \mathcal{R}_{\max}(\mathbf{h}^{fair}) > 0 \quad (43)$$

holds for all $\lambda, \mu \geq 0$. Notice that given $\lambda, \mu \geq 0$, this inequality holds for any $\xi^g, \xi^{l,k} > 0$. Let $\xi^g \rightarrow 0, \xi^{l,k} \rightarrow 0$, combining (42) and (43) gives that

$$\sum_{k \in [N]} \sum_{a \in \mathcal{A}} p_{a,k} \mathbb{E}_{X \sim \mathcal{P}_{a,k}^X} \left[\max_{y \in [m]} \left([\mathbf{M}^{\lambda, \mu}(a, k)]^\top \eta(X, a, k) \right)_y \right] > 0 \quad (44)$$

Therefore, the dual problem has lower bound

$$H(\lambda, \mu) \geq \xi^g \sum_{u_g \in \mathcal{U}_g} (\lambda_{u_g}^{(1)} + \lambda_{u_g}^{(2)}) + \sum_{k \in [N]} \xi^{l,k} \sum_{u_{l,k} \in \mathcal{U}_{l,k}} (\mu_{k, u_{l,k}}^{(1)} + \mu_{k, u_{l,k}}^{(2)}) \quad (45)$$

It presents that, as $\|\lambda\|_1 \rightarrow \infty$ or $\|\mu\|_1 \rightarrow \infty$, there must be $H(\lambda, \mu) \rightarrow \infty$, which conflicts with the dual problem $\min_{\lambda, \mu} H(\lambda, \mu)$. Hence, the optimal λ^*, μ^* of dual problem $\min_{\lambda, \mu} H(\lambda, \mu)$ must have bounded norms, denoting as $\|\lambda\|_1 \leq B_d, \|\mu\|_1 \leq B_d$.

Now we consider the differential of $H(\lambda, \mu)$. It is clear that $\{S_y = \{x \in \mathcal{X} : h_k(x, a) = y\}, y \in [m]\}$ constructs a partition of the feature space \mathcal{X} . Hence, for dual parameter $\lambda_{u_g}^{(1)}$, since the outer objective H is convex to λ and μ , by the additivity subgradients and Lemma B.4, the differential $\frac{\partial}{\partial \lambda_{u_g}^{(1)}} H(\lambda, \mu)$ can be formulated as

$$\frac{\partial}{\partial \lambda_{u_g}^{(1)}} H(\lambda, \mu) = \sum_{k \in [N]} \sum_{a \in \mathcal{A}} p_{a,k} \mathbb{E}_{X \sim \mathcal{P}_{a,k}^X} \left[\frac{\partial}{\partial \lambda_{u_g}^{(1)}} \max_{y \in [m]} \left([\mathbf{M}^{\lambda, \mu}(a, k)]^\top \eta(X, a, k) \right)_y \right] + \xi^g. \quad (46)$$

With a slight abuson of notation, let score function $f(x, a, k) = [\mathbf{M}^{\lambda, \mu}(a, k)]^\top \eta(x, a, k)$, by Lemma B.5, we have

$$\mathbb{E}_{X \sim \mathcal{P}_{a,k}^X} \left[\frac{\partial}{\partial \lambda_{u_g}^{(1)}} \max_{y \in [m]} \left([\mathbf{M}^{\lambda, \mu}(a, k)]^\top \eta(X, a, k) \right)_y \right] \quad (47)$$

$$= \sum_{y \in [m]} \int_{\{x: h_k^{\lambda, \mu}(x, a) = y\}} \left[\frac{\partial}{\partial \lambda_{u_g}^{(1)}} \max_{y \in [m]} \left([\mathbf{M}^{\lambda, \mu}(a, k)]^\top \eta(x, a, k) \right)_y \right] d\mathcal{P}_{a,k}^X(x) \quad (48)$$

$$= \frac{1}{p_{a,k}} \sum_{y \in [m]} \int_{\{x: h_k^{\lambda, \mu}(x, a) = y\}} \left[\text{conv} \left(\bigcup_{i \in \arg \max_i (f_i(x, a, k))} -[\eta(x, a, k)]^\top \mathbf{D}_{u_g}^{a,k} e_y \right) \right] d\mathcal{P}_{a,k}^X(x) \quad (49)$$

$$= \frac{1}{p_{a,k}} \sum_{y \in [m]} \left\{ \int_{\{x: f_y(x, a, k) \geq f_i(x, a, k), \forall i \neq y, i \in [m]\}} \left[-[\eta(x, a, k)]^\top \mathbf{D}_{u_g}^{a,k} e_y \right] d\mathcal{P}_{a,k}^X(x) \right. \\ \left. + \int_{B_y^t} \left[-b_t [\eta(x, a, k)]^\top \mathbf{D}_{u_g}^{a,k} (e_t - e_y) \right] d\mathcal{P}_{a,k}^X(x) \right\}, \quad (50)$$

where $B_y^t := \{\exists t \neq y, f_t(x, a, k) \geq f_i(x, a, k), \forall i \in [m]; f_t(x, a, k) = f_y(x, a, k)\}$ with $b_t \in [0, 1]$. Since the convex hull is a interval here, by Caratheodory's theorem, it can be characterized by

two point here (the initial point e_y and another point e_t in the convex hull). Without loss of generality, we assume the existence of one e_t such that $f_t(x, a, k) = f_y(x, a, k)$ here. We know that $f_t(x, a, k) - f_y(x, a, k) = [\eta(x, a, k)]^\top \mathbf{M}^{\lambda, \mu}(a, k)(e^t - e^y)$. With Asumption 1, we obtain that the measure of B_y^t is 0, unless the t -th and y -th column of $\mathbf{M}^{\lambda, \mu}(a, k)$ are equal. An effective simplification is to exclude all λ', μ' that cause $\mathbf{M}^{\lambda', \mu'}(a, k)(e^t - e^y) = 0$. Since we suppose that the non-zero columns of each $\mathbf{D}_u^{a, k}$ are distinct, the dual parameter $\lambda', \mu' \in S_{t, y}$, such that $\mathbf{M}^{\lambda', \mu'}(a, k)(e^t - e^y) = 0$, constructs the empty relative interior in the dual parameter space. By the convexity of the objective function, we have $\inf_{\lambda, \mu \notin S_{t, y}} H(\lambda, \mu) = \min_{\lambda, \mu} H(\lambda, \mu)$, due to the density of $(\lambda, \mu) \notin S_{t, y}$.

Overall, under the assumptions of the theorem, we have that B_y^t has a measure of zero. It follows that

$$\begin{aligned} \frac{\partial}{\partial \lambda_{u_g}^{(1)}} H(\lambda, \mu) &= \sum_{k \in [N]} \sum_{a \in \mathcal{A}} \sum_{y \in [m]} \int_{\{x: h_k^{\lambda, \mu}(x, a) = y\}} \left[- \left([\mathbf{D}_{u_g}^{a, k}]^\top \eta(x, a, k) \right)_y \right] d\mathcal{P}_{a, k}^X(x) + \xi^g \\ &= \sum_{k \in [N]} \sum_{a \in \mathcal{A}} \int_{\mathcal{X}} - [\eta(x, a, k)]^\top \mathbf{D}_{u_g}^{a, k} h_k^{\lambda, \mu}(x, a) d\mathcal{P}_{a, k}^X(x) + \xi^g \\ &= -\mathcal{D}_{u_g}^g(\mathbf{h}^{\lambda, \mu}) + \xi^g \end{aligned} \quad (51)$$

In a similar manner, we can derive

$$\begin{aligned} \frac{\partial}{\partial \lambda_{u_g}^{(2)}} H(\lambda, \mu) &= \sum_{k \in [N]} \sum_{a \in \mathcal{A}} \int_{\mathcal{X}} [\eta(x, a, k)]^\top \mathbf{D}_{u_g}^{a, k} h_k^{\lambda, \mu}(x, a) d\mathcal{P}_{a, k}^X(x) + \xi^g \\ &= \mathcal{D}_{u_g}^g(\mathbf{h}^{\lambda, \mu}) + \xi^g \end{aligned} \quad (52)$$

Considering paired optimal dual parameter $\lambda_{u_g}^{(i)*}, i = 1, 2$, by Lemma B.6, if $\lambda_{u_g}^{(1)*}, \lambda_{u_g}^{(2)*} > 0$, we have

$$\mathcal{D}_{u_g}^g(\mathbf{h}^{\lambda^*, \mu}) = -\xi^g, \mathcal{D}_{u_g}^g(\mathbf{h}^{\lambda^*, \mu}) = \xi^g,$$

which leads to a contradiction. If $\lambda_{u_g}^{(1)*} = 0, \lambda_{u_g}^{(2)*} = 0$, we have

$$\mathcal{D}_{u_g}^g(\mathbf{h}^{\lambda^*, \mu}) \geq -\xi^g, \mathcal{D}_{u_g}^g(\mathbf{h}^{\lambda^*, \mu}) \leq \xi^g.$$

If $\lambda_{u_g}^{(1)*} = 0, \lambda_{u_g}^{(2)*} > 0$, we have

$$\mathcal{D}_{u_g}^g(\mathbf{h}^{\lambda^*, \mu}) \geq -\xi^g, \mathcal{D}_{u_g}^g(\mathbf{h}^{\lambda^*, \mu}) = \xi^g.$$

If $\lambda_{u_g}^{(1)*} > 0, \lambda_{u_g}^{(2)*} = 0$, we have

$$\mathcal{D}_{u_g}^g(\mathbf{h}^{\lambda^*, \mu}) = -\xi^g, \mathcal{D}_{u_g}^g(\mathbf{h}^{\lambda^*, \mu}) \leq \xi^g.$$

Overall, we have shown that for all $u_g \in \mathcal{U}_g$, $|\mathcal{D}_{u_g}^g(\mathbf{h}^{\lambda^*, \mu})| \leq \xi^g$.

The local fairness guarantee also can be derived from the optimality of μ^* . The proof techniques are extremely similar to our proof with respect to λ^* . Hence, we omit the proof of the local fairness guarantee here. The result turns out that $|\mathcal{D}_{u_l, k}^{l, k}(\mathbf{h}^{\lambda^*, \mu^*})| \leq \xi^{l, k}, k \in [N]$.

The next step is to prove that the classifier $\mathbf{h}^{\lambda^*, \mu^*}$ is the optimal solution of the primal problem (1). From the proof above, we can obtain that, for $\forall u_g \in \mathcal{U}_g$,

$$(\lambda_{u_g}^{(1)} - \lambda_{u_g}^{(2)}) \mathcal{D}^g(\mathbf{h}^{\lambda^*, \mu^*}) - (\lambda_{u_g}^{(1)} + \lambda_{u_g}^{(2)}) \xi^g = 0,$$

which satisfies the optimality conditions for the dual solution of the constrained optimization problem. The same holds for the local fairness constraints $\mathcal{D}^{l, k}(\mathbf{h}^{\lambda^*, \mu^*})$. Consequently, the Lagrangian function equals to risk function when plugging in optimal classifier, $\mathcal{L}(h^{\lambda^*, \mu^*}, \lambda^*, \mu^*) = \mathcal{R}(h^{\lambda^*, \mu^*})$. For any other classifiers \mathbf{h}' that satisfies the global and local fairness constraints, denoting its corresponding dual parameter to maximize the outer problem as λ', μ' , it can be deduced that

$$\mathcal{L}(h^{\lambda^*, \mu^*}, \lambda^*, \mu^*) \leq \mathcal{L}(\mathbf{h}', \lambda', \mu') \leq \mathcal{R}(\mathbf{h}').$$

Therefore, we arrive at

$$\mathcal{R}(\mathbf{h}^{\lambda^*, \mu^*}) = \mathcal{L}(h^{\lambda^*, \mu^*}, \lambda^*, \mu^*) \leq \mathcal{R}(\mathbf{h}').$$

This completes the proof. \square

1249 B.7 Proof of Proposition 6

1250 **Note that** $\lambda \in \mathbb{R}_{\geq 0}^{2|\mathcal{U}_g|}$ and $\mu_k \in \mathbb{R}_{\geq 0}^{2|\mathcal{U}_{l,k}|}$, the operator $\|\cdot\|_1$ is linear in dual parameters' domain. We
 1251 can just write

$$\begin{aligned} \hat{H}'_k(\lambda, \mu_k) &:= \frac{1}{n_k} \sum_{i=1}^{n_k} \sigma_\beta([\hat{\mathbf{M}}^{\lambda, \mu}(a_i, k)]^\top \eta(x_i, a_i, k)) + \xi^g \sum_{u_g \in \mathcal{U}_g} (\lambda_{u_g}^{(1)} + \lambda_{u_g}^{(2)}) \\ &\quad + \frac{\xi^{l,k}}{\hat{p}_k} \sum_{u_{l,k} \in \mathcal{U}_{l,k}} (\mu_{k,u_{l,k}}^{(1)} + \mu_{k,u_{l,k}}^{(2)}), \end{aligned}$$

1252 where $\hat{\mathbf{M}}^{\lambda, \mu}(a, k) := \mathbf{I} - \frac{1}{\hat{p}_{a,k}} \left[\sum_{u_g \in \mathcal{U}_g} (\lambda_{u_g}^{(1)} - \lambda_{u_g}^{(2)}) \hat{\mathbf{D}}_{u_g}^{a,k} - \sum_{u_{l,k} \in \mathcal{U}_{l,k}} (\mu_{k,u_{l,k}}^{(1)} - \mu_{k,u_{l,k}}^{(2)}) \hat{\mathbf{D}}_{u_{l,k}}^{a,k} \right]$
 1253 and $\sigma_\beta(x) = \sum_{i=1}^m \frac{\exp(x_i/\beta)}{\sum_{j=1}^m \exp(x_j/\beta)} x_i$.

1254 **Convexity.** The $\hat{\mathbf{M}}^{\lambda, \mu}(a, k)$ is linear to λ and μ_k , and the soft-max operator is convex. Since the
 1255 composition of an affine mapping and a convex function preserves convexity, $\hat{H}'_k(\lambda, \mu_k)$ is convex to
 1256 λ and μ_k .

1257 **Smoothness.** Consider the soft-max weighted sum $\sigma_\beta(x) := \sum_{j=1}^m \frac{\exp(x_j/\beta)}{\sum_{\ell=1}^m \exp(x_\ell/\beta)} x_j$, and its
 1258 Hessian matrix is given by $H_\sigma(x) := \nabla^2 \sigma_\beta(x)$, $[H_\sigma(x)]_{i,j} = \frac{p_i}{\beta} \left[\left(2 + \frac{x_i - \bar{x}}{\beta} \right) \mathbb{I}(i=j) - p_j \left(2 + \frac{x_i + x_j - \bar{x}}{\beta} \right) \right]$. For $\forall i, j \in m$, if $\|x\|_1 \leq R$,
 1259

$$|[H_\sigma(x)]_{ij}| \leq \frac{1}{\beta} \left(\left(2 + \frac{2R}{\beta} \right) + \left(2 + \frac{4R}{\beta} \right) \right) = \frac{4\beta + 6R}{\beta^2}.$$

1260 Hence, its spectral norm is bounded,

$$\|H_\sigma(x)\|_2 \leq \|H_\sigma(x)\|_F \leq \left(\sum_{i,j \in [m]} [H_\sigma]_{ij}^2 \right)^{\frac{1}{2}} \leq m \frac{4\beta + 6R}{\beta^2}.$$

1261 Then, there exists a finite constant $L_\sigma := m \frac{4\beta + 6R}{\beta^2}$, such that $\|\nabla^2 \sigma_\beta(x)\|_2 \leq L_\sigma$.

For each sample $i = 1, \dots, n_k$, define the affine map

$$z_i(\lambda) := A_i \lambda + b_i, \quad [A_i]_{u_g}^{(j)} = \frac{3j-2}{\hat{p}_{a,k}} [\eta(x_i, a_i, k)]^\top \hat{\mathbf{D}}_{u_g}^{a_i,k} \quad \text{for } \lambda_{u_g}^{(j)}, j = 1, 2.$$

1262 Set $f_i(\lambda) := \sigma_\beta(z_i(\lambda))$. and let $f_i(\lambda) = \sigma_\beta(z_i(\lambda))$, $\sigma_\beta(x) = \sum_{j=1}^m \frac{e^{x_j/\beta}}{\sum_{\ell=1}^m e^{x_\ell/\beta}} x_j$,
 1263 By the chain rule and second-order derivatives, $\nabla_\lambda f_i(\lambda) = A_i^\top \nabla_x \sigma_\beta(z_i(\lambda))$, $\nabla_\lambda^2 f_i(\lambda) =$
 1264 $A_i^\top [\nabla^2 \sigma_\beta(z_i(\lambda))] A_i$. Hence, due to the boundedness of $\|\lambda\|_1$, the inside $z_i(\lambda)$ is bounded, set-
 1265 ting the upper bound as R for simplification here, $\|\nabla_\lambda^2 f_i(\lambda)\|_2 \leq \|A_i\|_2^2 \sup_x \|\nabla^2 \sigma_\beta(x)\|_2 =$
 1266 $\|A_i\|_2^2 L_\sigma$, showing f_i is $\|A_i\|_2^2 L_\sigma$ -smooth. The linear term in λ has zero Hessian. Therefore, since
 1267 the average of smooth functions is smooth with averaged constants, the function $\hat{H}'_k(\lambda, \mu_k)$ is L -
 1268 smooth in λ with $L = \frac{1}{n_k} \sum_{i=1}^{n_k} \|A_i\|_2^2 L_\sigma$. Following the similar proof procedure, we can obtain the
 1269 smoothness of $\hat{H}'_k(\lambda, \mu_k)$ to μ_k . \square

1270 B.8 Generalization Error For Post-processing Algorithm

1271 We begin by introducing some notations and simplifications, same as the proof of Theorem 9. Without
 1272 loss of generalization, let $n = n_1 = \dots = n_k$ present the sample number in local datasets. Denote
 1273 $p_{a|k} := \mathbb{P}(A = a | K = k)$, $p_{\min} := \min_{a \in \mathcal{A}, k \in [N]} p_{a|k}$. Assume $n_{\min} \geq 1$ denotes the sample size
 1274 of the sensitive group with the fewest observations across all clients.

1275 **Theorem 10.** If classifiers $\hat{\mathbf{h}}^* = (\hat{h}_1^*, \dots, \hat{h}_k^*)$ with dual parameters $(\hat{\lambda}^*, \hat{\mu}^*)$ form an optimal solution
 1276 of the empirical plug-in estimation of (7), denoting $\rho(n, \delta) = \sqrt{\frac{8|\mathcal{A}|m^2 \log(n+1)}{n}} + \sqrt{\frac{2 \log(m^2 |\mathcal{A}| N / \delta)}{n}}$,
 1277 $B_g = \max_{a \in \mathcal{A}, k \in [N]} \|\mathbf{D}_{u_g}^{a,k}\|_1$, $\hat{B}_g = \max_{a \in \mathcal{A}, k \in [N]} \|\hat{\mathbf{D}}_{u_g}^{a,k}\|_1$, $\Omega_n^g = \max_{a \in \mathcal{A}, k \in [N]} \|\mathbf{D}_{u_g}^{a,k} -$

1278 $\widehat{\mathbf{D}}_{u_g}^{a,k} \|_\infty$, and $B_{l,k} = \max_{a \in \mathcal{A}} \|\mathbf{D}_{u_{l,k}}^{a,k}\|_1$, $\widehat{B}_{l,k} = \max_{a \in \mathcal{A}} \|\widehat{\mathbf{D}}_{u_{l,k}}^{a,k}\|_1$, $\Omega_n^{l,k} = \max_{a \in \mathcal{A}} \|\mathbf{D}_{u_{l,k}}^{a,k} -$
1279 $\widehat{\mathbf{D}}_{u_{l,k}}^{a,k}\|_\infty$, $k \in [N]$.

1280 (1) Let $0 < \delta < 1$, suppose that $n > \frac{2|\mathcal{A}|N\widehat{B}_{l,k}}{p_{\min}\xi^g} + \frac{1}{2p_{\min}^2} \log \frac{1}{\delta}$, then with probability at least $1 - 2|\mathcal{A}|\delta$,

$$|\mathcal{D}^g(\widehat{\mathbf{h}}^*)| \leq \xi^g + \mathcal{O}(|\mathcal{A}|NB_g\rho(n, \delta/|\mathcal{A}||\mathcal{U}_g|)) + \Omega_n^g + \frac{|\mathcal{A}|N\widehat{B}_g}{n_{\min}}.$$

1281 (2) Let $0 < \delta < 1$, suppose that $n > \frac{2|\mathcal{A}|\widehat{B}_g}{p_{\min}\xi^{l,k}} + \frac{1}{2p_{\min}^2} \log \frac{1}{\delta}$, then with probability at least $1 - 2|\mathcal{A}|\delta$,

$$|\mathcal{D}^{l,k}(\widehat{\mathbf{h}}^*)| \leq \xi^{l,k} + \mathcal{O}(|\mathcal{A}|B_{l,k}\rho(n, N\delta/|\mathcal{A}||\mathcal{U}_g|)) + \Omega_n^{l,k} + \frac{|\mathcal{A}|\widehat{B}_{l,k}}{n_{\min}}, \quad k \in [N].$$

1282 The proof of Theorem 10 needs the following lemma.

1283 **Lemma B.7.** Let X_1, \dots, X_n be independent Bernoulli(p) random variables and define $S_n =$
1284 $\sum_{i=1}^n X_i$. Fix any $M \in (0, np)$ and confidence level $\delta \in (0, 1)$. If the sample size satisfies $n \geq$
1285 $\frac{2M}{p} + \frac{1}{2p^2} \log \frac{1}{\delta}$, then we have $\mathbb{P}(S_n > M) \geq 1 - \delta$.

Proof of Lemma B.7. By Hoeffding's inequality, for any $t > 0$, $\mathbb{P}(S_n - \mathbb{E}[S_n] \leq -t) \leq \exp\left(-\frac{2t^2}{n}\right)$.
Since $\mathbb{E}[S_n] = np$, set $t = np - M$. Then

$$\mathbb{P}(S_n \leq M) = \mathbb{P}(S_n - np \leq -(np - M)) \leq \exp\left(-\frac{2(np - M)^2}{n}\right).$$

To guarantee $\mathbb{P}(S_n \leq M) \leq \delta$, it suffices that $\frac{2(np - M)^2}{n} \geq \log \frac{1}{\delta}$. Substitute $n = \frac{2M}{p} + \frac{1}{2p^2} \log \frac{1}{\delta}$.
Then $np - M = \left(\frac{2M}{p} + \frac{1}{2p^2} \log \frac{1}{\delta}\right)p - M = M + \frac{1}{2p} \log \frac{1}{\delta}$, and one can check

$$\frac{2(np - M)^2}{n} = \frac{2\left(M + \frac{1}{2p} \log \frac{1}{\delta}\right)^2}{\frac{2M}{p} + \frac{1}{2p^2} \log \frac{1}{\delta}} \geq \log \frac{1}{\delta}.$$

1286 Hence $\mathbb{P}(S_n \leq M) \leq \delta$, i.e. $\mathbb{P}(S_n > M) \geq 1 - \delta$. □

1287 B.8.1 Proof of Theorem 10

1288 We first consider the generalization error of the fairness constraints. Without loss of generalization,
1289 here we only prove the generalization error for global fairness constraints and corresponding parameter
1290 λ . The proof technique for local fairness constraints and corresponding parameter μ is extremely
1291 similar to that for global fairness constraints.

1292 We know that the personalized attribute-aware empirical classifier can be written as

$$\widehat{h}_k^{\widehat{\lambda}^*, \widehat{\mu}^*}(x, a) := \arg \max_{y \in [m]} [\widehat{\mathbf{M}}^{\widehat{\lambda}^*, \widehat{\mu}^*}(a, k)]^\top \widehat{\eta}(x, a, k) \quad (53)$$

As h depends on the Bayes score function η , we can consider the input as $(\eta_{k,i} :=$
 $\eta(x_{k,i}, a_{k,i}, k), a_{k,i}, y_{k,i})$. Let $\ell_{i,j}^{a',k}(\eta, a, y; h) = \mathbb{I}(y = i \wedge h(\eta) = j \wedge a = a')$. Then we
have $\mathbf{C}_{i,j}^{a',k}(h) = \mathbb{E}[\ell_{i,j}^{a',k}(\eta, y, a; h)]$ and $\widehat{\mathbf{C}}_{i,j}^{a',k}(h) = \frac{1}{n} \sum_{z=1}^n \ell_{i,j}^{a',k}(\eta_{k,z}, a_{k,z}, y_{k,z})$. Then we turn
to consider the VC dimension of the function class $\mathcal{H}_{i,j,a'} := \{h : (x, a, y) \rightarrow \mathbb{I}(y = i \wedge h(\eta) =$
 $j \wedge a = a')\}$. Thanks to the classifier's specific structural form (53), we can directly state an explicit
upper bound on its VC dimension: for given class j ,

$$\widehat{h}_k(x, a) = j \Leftrightarrow [\eta(x, a, k)]^\top \left([\widehat{\mathbf{M}}^{\widehat{\lambda}^*, \widehat{\mu}^*}(a, k)]_{:,j} - [\widehat{\mathbf{M}}^{\widehat{\lambda}^*, \widehat{\mu}^*}(a, k)]_{:,i} \right) \geq 0, \quad \forall i \neq j \in [m],$$

1293 which can be regarded as the intersection of $m - 1$ half-spaces given $\eta_{k,i}, a_{k,i}$. A single halfspace
1294 function class can be viewed as the class of linear classifiers, possessing a VC dimension of m . By the
1295 additive property of VC dimension, for function classes $\{\mathcal{G}_i\}_{i=1}^m$, $VC(\bigwedge_{i=1}^m \mathcal{G}_i) \leq \sum_{i=1}^m VC(\mathcal{G}_i)$,

1296 the function class $\mathcal{H}_{i,j,a'}$ has VC dimension at most $\mathcal{O}(|\mathcal{A}|m^2)$. By taking a union bound in the
 1297 Lemma B.3, we have that with probability at least $1 - \delta$,

$$\sup_{h \in \mathcal{F}_{\mathcal{H}}} \max_{k \in [N]} \max_{a \in \mathcal{A}} \|\mathbf{C}^{a,k}(h) - \widehat{\mathbf{C}}^{a,k}(h)\|_{\infty} \leq \mathcal{O}\left(\sqrt{\frac{8|\mathcal{A}|m^2 \log(n+1)}{n}} + \sqrt{\frac{2 \log(m^2|\mathcal{A}|N/\delta)}{n}}\right) \\ := \mathcal{O}(\rho(n, \delta)).$$

1298 Hence, for the global fairness constraints $\mathcal{D}_{u_g}^g$ with the empirical optimal solution $\widehat{\mathbf{h}}^*$, by the
 1299 generalization bound in (29), we have that,

$$\mathcal{D}_{u_g}^g(\widehat{\mathbf{h}}) - \widehat{\mathcal{D}}_{u_g}^g(\widehat{\mathbf{h}}) \leq \mathcal{O}(\rho(n, \delta/|\mathcal{A}|))|\mathcal{A}|NB_g + \Omega_n^g.$$

1300 Now we consider the bound on empirical $\widehat{\mathcal{D}}_{u_g}^g(\widehat{\mathbf{h}}^*)$. The empirical optimal dual parameter $\widehat{\lambda}^*$ and $\widehat{\mu}^*$
 1301 are obtained by the empirical dual function:

$$\widehat{H}(\lambda, \mu) = \sum_{k \in [N]} \sum_{a \in \mathcal{A}} \widehat{p}_{a,k} \sum_{i=1}^{n_{a,k}} \left[\max_{y \in [m]} \left([\widehat{\mathbf{M}}^{\lambda, \mu}(a, k)]^{\top} \widehat{\eta}(x_i, a, k) \right)_y \right] + \xi^g \sum_{u_g \in \mathcal{U}_g} (\lambda_{u_g}^{(1)} + \lambda_{u_g}^{(2)}) \\ + \sum_{k \in [N]} \xi^{l,k} \sum_{u_{l,k} \in \mathcal{U}_{l,k}} (\mu_{k, u_{l,k}}^{(1)} + \mu_{k, u_{l,k}}^{(2)}). \quad (54)$$

1302 This representation is fully consistent with that given in (8) restricting to group- a observations within
 1303 the k -th client's data, where $n_{a,k}$ denotes the sample number of group a in client k . Considering the
 1304 subgradient of the empirical dual function w.r.t. $\lambda_{u_g}^{(1)}$, by the additivity of subgradient,

$$\frac{\partial}{\partial \lambda_{u_g}^{(1)}} \widehat{H}(\lambda, \mu) = \sum_{k \in [N]} \sum_{a \in \mathcal{A}} \widehat{p}_{a,k} \frac{1}{n_{a,k}} \sum_{i=1}^{n_{a,k}} \left[\frac{\partial}{\partial \lambda_{u_g}^{(1)}} \max_{y \in [m]} \left([\widehat{\mathbf{M}}^{\lambda, \mu}(a, k)]^{\top} \widehat{\eta}(x_i, a, k) \right)_y \right] + \xi^g \quad (55)$$

1305 Denoting empirical score function $\widehat{f}(x, a, k) = [\widehat{\mathbf{M}}^{\lambda, \mu}(a, k)]^{\top} \widehat{\eta}(x, a, k)$, by Lemma B.5, we have

$$\sum_{i=1}^{n_{a,k}} \left[\frac{\partial}{\partial \lambda_{u_g}^{(1)}} \max_{y \in [m]} \left([\widehat{\mathbf{M}}^{\lambda, \mu}(a, k)]^{\top} \widehat{\eta}(x_i, a, k) \right)_y \right] \\ = \sum_{i=1}^{n_{a,k}} \sum_{y \in [m]} \mathbb{I}(\widehat{h}_k(x_i, a) = y) \left[\frac{\partial}{\partial \lambda_{u_g}^{(1)}} \max_{y \in [m]} \left([\widehat{\mathbf{M}}^{\lambda, \mu}(a, k)]^{\top} \widehat{\eta}(x_i, a, k) \right)_y \right] + \xi^g \\ = \frac{1}{\widehat{p}_{a,k}} \sum_{i=1}^{n_{a,k}} \sum_{y \in [m]} \mathbb{I}(\widehat{h}_k(x_i, a) = y) \left[\text{conv} \left(\bigcup_{i \in \arg \max_{j \in [m]} \widehat{f}_j(x, a, k)} -[\widehat{\eta}(x_i, a, k)]^{\top} \widehat{\mathbf{D}}_{u_g}^{a,k} e_i \right) \right] \\ = \frac{1}{\widehat{p}_{a,k}} \sum_{i=1}^{n_{a,k}} \sum_{y \in [m]} \left\{ -\mathbb{I} \left(\widehat{f}_y(x_i, a, k) > \widehat{f}_j(x_i, a, k), \forall j \neq y, j \in [m] \right) [\widehat{\eta}(x_i, a, k)]^{\top} \widehat{\mathbf{D}}_{u_g}^{a,k} e_y \right. \\ \left. + \mathbb{I} \left(x_i \in B_y^t \right) \left[-b_t [\widehat{\eta}(x, a, k)]^{\top} \widehat{\mathbf{D}}_{u_g}^{a,k} (e_t - e_y) \right] \right\}$$

1306 where $B_y^t := \{x : \exists t \neq y, \widehat{f}_t(x, a, k) \geq \widehat{f}_i(x, a, k), \forall i \in [m]; \widehat{f}_t(x, a, k) = \widehat{f}_y(x, a, k)\}$ and
 1307 $b_t \in [0, 1]$. According to Carathéodory's theorem, the subgradient interval can still be represented by
 1308 two points. According to our assumption, the plug-in estimator $\widehat{\eta}$ still meet the continuity assumption
 1309 and we exclude singular λ', μ' . Therefore, we know that

$$\mathbb{P} \left(\sum_{i=1}^{n_{a,k}} \mathbb{I} \left(\exists t \neq y, \widehat{f}_t(x_i, a, k) \geq \widehat{f}_i(x_i, a, k), \forall i \in [m]; \widehat{f}_t(x_i, a, k) = \widehat{f}_y(x_i, a, k) \right) \leq 1 \right) = 1 \quad (57)$$

1310 Hence, the subgradient falls into an interval. Since $[\hat{\eta}(x_i, a, k)]^\top \hat{\mathbf{D}}_{u_g}^{a,k} e_t \leq \|[\hat{\eta}(x_i, a, k)]^\top \hat{\mathbf{D}}_{u_g}^{a,k}\|_1 \leq$
 1311 B_g , denoting $n_{\min} := \min_{a \in \mathcal{A}, k \in [N]} n_{a,k}$ and \hat{B}_g , we have that

$$\frac{\partial}{\partial \lambda_{u_g}^{(1)}} \hat{H}(\lambda, \mu) \leq \sum_{k \in [N]} \sum_{a \in \mathcal{A}} \frac{1}{n_{a,k}} \sum_{i=1}^{n_{a,k}} \left[-[\hat{\eta}(x_i, a, k)]^\top \hat{\mathbf{D}}_{u_g}^{a,k} \hat{h}_k(x_i, a) \right] + \xi^g \quad (58)$$

$$+ \sum_{k \in [N]} \sum_{a \in \mathcal{A}} \frac{1}{n_{a,k}} \hat{B}_g \quad (59)$$

$$\leq -\hat{\mathcal{D}}_{u_g}^g(\hat{\mathbf{h}}) + \xi^g + \frac{|\mathcal{A}|N\hat{B}_g}{n_{\min}} \quad (60)$$

1312 On the other hand,

$$\frac{\partial}{\partial \lambda_{u_g}^{(1)}} \hat{H}(\lambda, \mu) \geq \sum_{k \in [N]} \sum_{a \in \mathcal{A}} \frac{1}{n_{a,k}} \sum_{i=1}^{n_{a,k}} \left[-[\hat{\eta}(x_i, a, k)]^\top \hat{\mathbf{D}}_{u_g}^{a,k} \hat{h}_k(x_i, a) \right] + \xi^g \quad (61)$$

$$- \sum_{k \in [N]} \sum_{a \in \mathcal{A}} \frac{1}{n_{a,k}} \hat{B}_g \quad (62)$$

$$= -\hat{\mathcal{D}}_{u_g}^g(\hat{\mathbf{h}}) + \xi^g - \frac{|\mathcal{A}|N\hat{B}_g}{n_{\min}} \quad (63)$$

1313 Hence, we obtain that

$$\frac{\partial}{\partial \lambda_{u_g}^{(1)}} \hat{H}(\lambda, \mu) - \xi^g \subset \left[-\hat{\mathcal{D}}_{u_g}^g(\hat{\mathbf{h}}^*) - \frac{|\mathcal{A}|N\hat{B}_g}{n_{\min}}, -\hat{\mathcal{D}}_{u_g}^g(\hat{\mathbf{h}}^*) + \frac{|\mathcal{A}|N\hat{B}_g}{n_{\min}} \right].$$

1314 In a similar manner, we can derive the range of subgradient for $\lambda_{u_g}^{(2)}$,

$$\frac{\partial}{\partial \lambda_{u_g}^{(2)}} \hat{H}(\lambda, \mu) - \xi^g \subset \left[\hat{\mathcal{D}}_{u_g}^g(\hat{\mathbf{h}}^*) - \frac{|\mathcal{A}|N\hat{B}_g}{n_{\min}}, \hat{\mathcal{D}}_{u_g}^g(\hat{\mathbf{h}}^*) + \frac{|\mathcal{A}|N\hat{B}_g}{n_{\min}} \right].$$

Since we assume that $n > \frac{2|\mathcal{A}|N\hat{B}_{t,k}}{p_{\min}\xi^g} + \frac{1}{2p_{\min}^2} \log \frac{1}{\delta}$, by Lemma B.7, we have that with probability at least $1 - |\mathcal{A}|\delta$,

$$n_{\min} \geq \frac{|\mathcal{A}|N\hat{B}_g}{\xi^g} \Leftrightarrow \xi^g \geq \frac{|\mathcal{A}|N\hat{B}_g}{n_{\min}}.$$

1315 Consider the optimality of $\hat{\lambda}_{u_g}^{(1)}, \hat{\lambda}_{u_g}^{(2)}$, by Lemma B.6, if $\hat{\lambda}_{u_g}^{(1)} > 0, \hat{\lambda}_{u_g}^{(2)} > 0$, we have $0 \in$
 1316 $\frac{\partial}{\partial \lambda_{u_g}^{(1)}} \hat{H}(\hat{\lambda}^*, \mu), 0 \in \frac{\partial}{\partial \lambda_{u_g}^{(2)}} \hat{H}(\hat{\lambda}^*, \mu)$. Thus,

$$|\hat{\mathcal{D}}_{u_g}^g(\hat{\mathbf{h}}^*) - \xi^g| \leq \frac{|\mathcal{A}|N\hat{B}_g}{n_{\min}}$$

$$|\hat{\mathcal{D}}_{u_g}^g(\hat{\mathbf{h}}^*) + \xi^g| \leq \frac{|\mathcal{A}|N\hat{B}_g}{n_{\min}},$$

1317 which leads to a contradiction. For other cases, such as $\hat{\lambda}_{u_g}^{(1)} = \hat{\lambda}_{u_g}^{(2)} = 0; \hat{\lambda}_{u_g}^{(1)} > 0, \hat{\lambda}_{u_g}^{(2)} = 0$, and
 1318 $\hat{\lambda}_{u_g}^{(1)} = 0, \hat{\lambda}_{u_g}^{(2)} > 0$, as discussed in the proof of Theorem 5, it turns out that

$$|\hat{\mathcal{D}}_{u_g}^g(\hat{\mathbf{h}}^*)| \leq \xi^g + \frac{|\mathcal{A}|N\hat{B}_g}{n_{\min}}.$$

1319 By taking a union bound, we obtain that with probability at least $1 - 2|\mathcal{A}|\delta$,

$$|\mathcal{D}^g(\hat{\mathbf{h}}^*)| \leq \xi^g + \mathcal{O}(\rho(n, \delta/|\mathcal{A}||\mathcal{U}_g|))|\mathcal{A}|NB_g + \Omega_n^g + \frac{|\mathcal{A}|N\hat{B}_g}{n_{\min}}.$$

1320 For local fairness constraints $\mathcal{D}^{l,k}(\hat{\mathbf{h}}^*)$, following the same proof procedures, we arrive at that with
 1321 probability at least $1 - 2|\mathcal{A}|\delta$,

$$|\mathcal{D}^{l,k}(\hat{\mathbf{h}}^*)| \leq \xi^{l,k} + \mathcal{O}(\rho(n, N\delta/|\mathcal{A}||\mathcal{U}_g|))|\mathcal{A}|B_{l,k} + \Omega_n^{l,k} + \frac{|\mathcal{A}|\hat{B}_{l,k}}{n_{\min}}.$$

1322

□

C Additional Datasets and Experimental Setting

C.1 Datasets and Experimental Details

C.1.1 Datasets

- The **Compas** dataset [20] comprises 6,172 criminal defendants from Broward County, Florida, between 2013 and 2014, with the task of predicting whether a defendant will recidivate within two years of their initial risk assessment. We consider the race of each individual as the sensitive attribute and train a logistic classifier as our prediction model.
- The **Adult** dataset [4] comprises more than 45000 samples based on 1994 U.S. census data, where the task is to predict whether the annual income of an individual is above \$50,000. We consider the gender of each individual as the sensitive attribute and train the logistic regression as the classification model.
- The **ENEM** dataset [36] contains about 1.4 million samples from Brazilian college entrance exam scores along with student demographic information. We follow [3] to quantized the exam score into 2 or 5 classes as label, and consider race as sensitive attribute. As [3] used a random subset of 50K samples, we instead sample 100K data points to construct our federated dataset. We train multilayer perceptron (MLP) as the classification model.
- The **CelebA** dataset [83] is a facial image dataset consists of about 200k instances with 40 binary attribute annotations. We identify the binary feature *smile* as target attributes which aims to predict whether the individuals in the images exhibit a smiling expression. The *race* of individuals is chosen as sensitive attribute. We train Resnet18 [34] on CelebA as the classification model.

The determination of sensitive attributes and labels on three datasets has been verified significant in previous research [3, 31].

C.1.2 Baselines

We compare the performance of FedFACT with traditional **FedAvg** [50] and five SOTA methods tailored for calibrating global and local fairness in FL, namely **FairFed** [29], **FedFB** [81], **FCFL** [17], **praFFL** [77], and the method in [23], denoted as **Cost** in our experiments.

- **FedAvg** serves as a core Federated Learning model and provides the baseline for our experiments. It works by computing updates on each client’s local dataset and subsequently aggregating these updates on a central server via averaging.
- **FairFed** introduces an approach to adaptively adjust the aggregation weights of different clients based on their local fairness metric to train federated model with global fairness guarantee.
- **FedFB** presents a FairBatch-based approach [60] to compute the coefficients of FairBatch parameters on the server. This method integrates global reweighting for each client into the FedAvg framework to fulfill fairness objectives.
- **FCFL** proposed a two-stage optimization to solve a multi-objective optimization with fairness constraints. The prediction loss at each local client is treated as an objective, and FCFL maximize the worst-performing client while considering fairness constraints by optimizing a surrogate maximum function involving all objectives.
- **praFFL** proposed a preference-aware federated learning scheme that integrates client-specific preference vectors into both the shared and personalized model components via a hypernetwork. It is theoretically proven to linearly converge to Pareto-optimal personalized models for each client’s preference.
- **[23]** proposed a convex-programming-based post-processing framework that characterizes and enforces the minimum accuracy loss required to satisfy specified levels of both local and global fairness constraints in multi-class federated learning by approximating the region under the ROC hypersurface with a simplex and solving a linear program, denoted as **Cost** in our experiments.

Meanwhile, we adapt FedFACT to focus solely on global or local fairness in FL, denoted as FedFACT_g and FedFACT_l . $\text{FedFACT}_{g\&l}$ indicates the algorithm simultaneously achieving global and local fairness. The FedFACT (In) presents the in-processing method and FedFACT (Post) presents the post-processing method.

C.1.3 Parameter Settings

We provide hyperparameter selection ranges for each model in Table 4. For all other hyperparameters, we follow the codes provided by authors and retain their default parameter settings.

Table 4: Hyperparameter Selection Ranges

Model	Hyperparameter	Ranges
General	Learning rate	{0.0001, 0.001, 0.003, 0.005, 0.01, 0.03, 0.05}
	Global round	{20, 30, 50, 80}
	Local round	{10, 20, 30, 50}
	Local batch size	{128, 256, 512}
	Hidden layer	{16, 32, 64}
	Optimizer	{Adam, SGD}
FedFB	Step size (α)	{0.005, 0.01, 0.05, 0.3}
FairFed	Fairness budget (β)	{0.01, 0.05, 0.5, 1}
	Local debiasing (α)	{0.005, 0.01, 0.05}
FCFL	Fairness constraint (ϵ)	{0.01, 0.03, 0.05, 0.07}
praFFL	Diversity (τ_p)	{10, 15, 20}
FedFACT (In)	Classifier number	1
	w_k^t learning rate (η_w)	{0.03, 0.3}
	Dual parameter bound	5
FedFACT (Post)	Temperature β	0.1
	Dual parameter bound	5

For the fairness-control parameters, e.g., the parameter λ in praFFL [77] and the global and local fairness constraints in Cost [23], we impose stringent fairness requirements on the model in our overall comparative experiments, and we adjust the parameters governing the fairness metrics in the Pareto-curve experiments.

C.1.4 Experiments Compute Resources

We conducted our experiments on a GPU server equipped with 8 CPUs and two NVIDIA RTX 4090s (24G).

C.2 Discussion about FedFACT and LoGoFair [82]

LogoFair [82] is designed for binary-classification in federated learning under both global and local fairness constraints, seeking the Bayes-optimal classifier. By deriving a closed-form solution for the fair Bayes classifier, LogoFair reformulates the post-processing fairness adjustment as a bilevel optimization problem jointly solved by the server and clients, which is an approach conceptually analogous to our post-processing framework. In binary classification, FedFact and LogoFair both target Bayes-optimal classifiers under constraints disparity metrics expressed in linear form. Theoretically, for an identical fairness metric, our Bayes-optimal fair classifier characterization covers that of LogoFair. Consequently, we refrain from performing a comparative evaluation of the two approaches.

Our method differs by defining the loss at the client level, thereby achieving lower estimation error than the local group-specific objective in [82]. Crucially, by formulating the post-processing model over the probabilistic simplex instead of restricting outputs to the unit interval $[0, 1]$ in the binary case, our framework achieves enhanced scalability and naturally adaptable to multi-group, multiclass settings.

Note that, whether for binary or multiclass settings, our implementation of FedFACT is based on calibrating confusion matrices over the multi-dimensional probabilistic simplex.

1399 C.3 Heterogeneous Split of Client Distribution

1400 We propose a partitioning method that introduces heterogeneous correlations between the sensitive
1401 attribute A and label Y , thereby further elucidating the trade-off between global fairness and local
1402 fairness [30].

Heterogeneous Split. We assume a dataset D of n samples, each with a binary attribute A and a binary label Y . We denote by $n_{ij} = |\{x_\ell, a_\ell, y_\ell : (a_\ell = i, y_\ell = j)\}|$ the number of samples in joint class (i, j) for $i, j \in \{0, 1\}$. Our goal is to partition D into N disjoint subsets (one per client) such that in client $k \in [N]$, the correlation between A and Y is controlled by a target parameter $\gamma_k \in [a, b] \subseteq [0, 1]$. To achieve this, we first assign each client k a weight γ_k

$$w_k^{(i,j)} = \begin{cases} \gamma_k, & (i, j) \in \{(0, 0), (1, 1)\}, \\ 1 - \gamma_k, & (i, j) \in \{(1, 0), (0, 1)\}. \end{cases}$$

1403 Then for each joint class (i, j) we compute the total weight $W^{(i,j)} = \sum_{k=1}^n w_k^{(i,j)}$ and assign to
1404 client k a preliminary count $c_k^{(i,j)} = \lfloor (w_k^{(i,j)} / W^{(i,j)}) n_{ij} \rfloor$. Any remaining samples are distributed
1405 one by one to the clients with the largest fractional remainders, so that $\sum_{k=1}^N c_k^{(i,j)} = n_{ij}$. Finally,
1406 for each class (i, j) we shuffle its n_{ij} sample indices and slice them into blocks of size $c_k^{(i,j)}$. Client
1407 k then collects all its four blocks across (i, j) , yielding a partition that in expectation realizes the
1408 desired within-client correlation γ_k between A and Y .

1409 This approach can be regarded as a generalization of the synergy-level-based heterogeneous split
1410 in [30] to the multi-client setting, where the A - Y correlation for each client is governed by a parameter
1411 randomly drawn from $[a, b] \subseteq [0, 1]$, thereby yielding a more pronounced balance between global
1412 fairness and local fairness. Throughout the experimental evaluation, we set $\gamma_k \in [0.2, 0.8]$ to
1413 guarantee that every client has a sufficient number of sensitive group samples to assess local fairness.

1414 D Detailed Experiments Results

1415 D.1 Comparison Result and binary EO criterion

1416 **Parato Curves of DP.** We have already presented the numerical comparison between our proposed
1417 method and the baselines in the main text; here, we report the Pareto curves illustrating the trade-off
1418 between global fairness and accuracy. More precisely, we compare the trade-off between accuracy
1419 and the global fairness measure, as well as the trade-off between accuracy and the local fairness
1420 measure, as a function of the fairness constraint.

1421 The Pareto curve for the global DP criterion is shown in Figure 1 where the horizontal axis denotes
1422 accuracy and the vertical axis represents the fairness metric. Consequently, models located closer to
1423 the upper-right corner exhibit superior accuracy-fairness trade-offs. As illustrated in Figure 1, our
1424 method outperforms all existing state-of-the-art approaches when comparing accuracy against either
1425 global fairness in isolation.

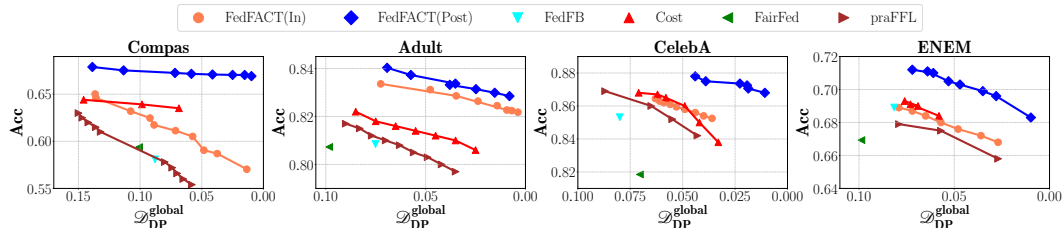


Figure 1: The Pareto frontier on Compas, Adult, CelebA and ENEM datasets. The curve closer to the upper right corner indicates a better trade-off between accuracy and fairness.

1426 This result not only demonstrates that our model achieves a more favorable accuracy-fairness balance
1427 but also highlights its controllability: by tuning the fairness constraints, one can satisfy diverse
1428 fairness requirements.

1429 **Parato Curves of EO.** In Figure 2, we illustrate the Pareto curve for the Equalized Odds (EO)
 1430 criterion-accuracy. Because EO enforces tighter constraints than DP, precise adherence in a federated
 1431 context requires large per-group sample counts at each client. Hence, we also compare the global
 1432 EO here. Our framework still exceeds all state-of-the-art baselines in trading off accuracy against
 1433 fairness.

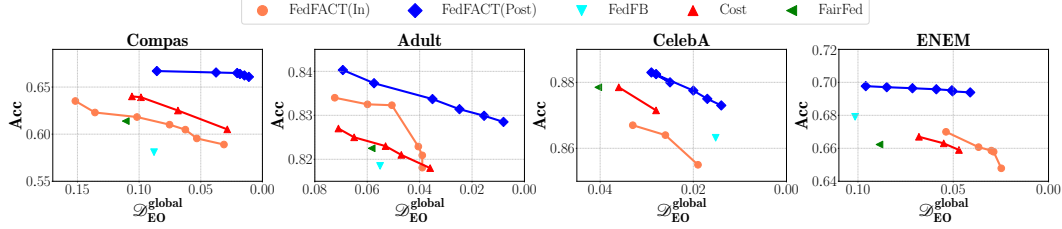


Figure 2: The Pareto frontier on Compas, Adult, CelebA and ENEM datasets. The curve closer to the upper right corner indicates a better trade-off between accuracy and fairness.

1434 D.2 Result for multi-class classification

1435 **Multi-Class fair datasets.** We illustrate how FedFACT performs on multi-class prediction using
 1436 CelebA and ENEM. For CelebA, with 'Gender' still serving as the sensitive attribute, We employ the
 1437 binary attributes "Smile" and "Big_Nose" to construct a multiclass task by mapping their joint values
 1438 $\{0, 1\} \times \{0, 1\}$ onto a four-class label set $\{0, 1, 2, 3\}$, thereby formulating a multiclass classification
 1439 problem on the CelebA dataset. These attributes are commonly used in centralized machine learning
 1440 literature [12, 84] to construct fairness-aware classification tasks. For ENEM, we follow [3] to
 1441 quantize the Humanities exam score to 5 classes. In order to guarantee adequate per-group sample
 1442 sizes at each client in heterogeneous settings for fairness evaluation (or some clients only hold less
 1443 than 10 samples for specific group under heterogeneous partitioning), we adopt the four race labels
 1444 "Branca," "Preta," "Parda," and "Amarela" from the Race attribute as the sensitive groups. These
 1445 datasets are partitioned into five clients under a heterogeneous split with $\gamma = 1$.

1446 **Evaluation.** In terms of baselines, only the Cost [23] algorithm is theoretically applicable to fairness
 1447 optimization in multiclass federated learning scenarios. However, their experiments and code are
 1448 limited to binary classification, and have already been used as baselines for comparison with our
 1449 method. Consequently, we focus exclusively on reporting FedFACT's performance along with
 1450 FedAvg in multiclass fairness, establishing it as a pioneering approach in this setting. In Figure 3, we
 plot the global-local fairness-accuracy trade-off of FedFACT in the multi-class classification task.

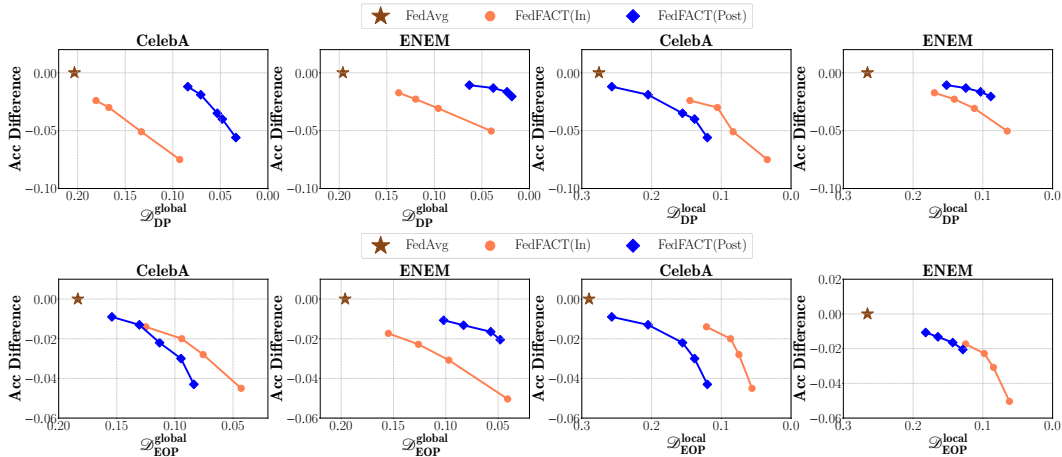


Figure 3: Multi-Class Fair Classification Results. Multiclass fairness calibration experiments in federated learning on the CelebA and ENEM datasets. The top panels depict global and local multiclass Demographic Parity (DP) results, while the bottom panels show global and local multiclass Equal Opportunity (EOP) outcomes.

Multi-Class Results. The outcomes of the multiclass experiments do not parallel those in the binary setting, where post-processing methods vastly outperform alternatives; instead, performance is comparatively lower. This can be attributed to the paucity of local samples for fairness evaluation at individual clients, which causes post-processing under joint global and local fairness constraints to incur significant generalization error and thus fail to precisely enforce local fairness (e.g. multiclass local EOP experiments in Figure 3). Under these conditions, the in-training approach, leveraging globally aggregated data, offers superior fairness calibration, thereby underscoring the complementarity of the two methods we introduce.

D.3 Additional Experiments for Adjusting Accuracy-Fairness Trade-Off

In Table 5, we present additional experiments on the Compas and Adult datasets under the heterogeneous split to illustrate the adjustment of the accuracy-fairness trade-off. Compared to the results in the main text, this partitioning yields a more pronounced trade-off between global and local fairness.

Table 5: Additional Accuracy-Fairness Balance.

Dataset (ξ^g, ξ^l)	Compas (In-)			Adult (In-)			Compas (Post-)			Adult (Post-)		
	Acc	\mathcal{J}^{global}	\mathcal{J}^{local}	Acc	\mathcal{J}^{global}	\mathcal{J}^{local}	Acc	\mathcal{J}^{global}	\mathcal{J}^{local}	Acc	\mathcal{J}^{global}	\mathcal{J}^{local}
(0.00,0.00)	60.22	0.0404	0.0745	80.99	0.0021	0.0407	64.56	0.0083	0.0075	81.25	0.0139	0.0275
(0.02,0.00)	60.61	0.0436	0.0734	81.04	0.0021	0.0423	64.78	0.0091	0.0099	81.56	0.0146	0.0285
(0.04,0.00)	60.90	0.0490	0.0737	81.09	0.0039	0.0446	65.04	0.0123	0.0099	81.62	0.0147	0.0285
(0.00,0.02)	60.80	0.0499	0.0744	81.18	0.0046	0.0411	64.94	0.0146	0.0214	81.82	0.0238	0.0381
(0.02,0.02)	61.03	0.0503	0.0726	81.64	0.0315	0.0463	65.12	0.0311	0.0306	82.04	0.0240	0.0381
(0.04,0.02)	61.32	0.0555	0.0774	81.65	0.0318	0.0467	65.57	0.0378	0.0371	82.16	0.0257	0.0397
(0.00,0.04)	61.18	0.0581	0.0804	81.31	0.0053	0.0444	65.16	0.0294	0.0517	82.46	0.0350	0.0492
(0.02,0.04)	61.39	0.0644	0.0753	81.67	0.0177	0.0452	65.16	0.0412	0.0419	82.49	0.0346	0.0497
(0.04,0.04)	62.39	0.0878	0.0966	82.14	0.0486	0.0497	65.82	0.0507	0.0574	82.63	0.0350	0.0518

Note that the gap between the imposed constraints and the observed fairness metrics stems from the **inevitable generalization error** incurred with finite local samples. Consequently, global fairness exhibits greater controllability than local fairness. In practice, FedFACT remains capable of tuning the accuracy-fairness balance according to the specified fairness constraints, highlighting the controllability inherent in our approach.

D.4 Hyper-Parameter Experiments

In this subsection, we examine the impact of the number of classifiers in the in-processing method. Specifically, we incrementally increase the size of the weighted ensemble—from using only the most recently trained classifier up to including the ten preceding classifiers. Let N_h represent the number of classifiers comprising the weighted ensemble. As reported in Table 6, we observe that augmenting the ensemble with multiple classifiers yields negligible improvements and can even degrade performance when earlier classifiers have not been fully trained. Consequently, in light of these empirical findings, all in-processing experiments in this work utilize only the single most recently obtained classifier.

Table 6: Hyper-Parameter Experimental Results.

N_h	Compas			Adult			CelebA			ENEM		
	Acc	\mathcal{J}^{global}	\mathcal{J}^{local}	Acc	\mathcal{J}^{global}	\mathcal{J}^{local}	Acc	\mathcal{J}^{global}	\mathcal{J}^{local}	Acc	\mathcal{J}^{global}	\mathcal{J}^{local}
1	61.17	0.0407	0.0732	82.04	0.0014	0.0401	86.15	0.0382	0.0473	68.33	0.0493	0.0487
2	61.29	0.0408	0.0731	81.24	0.0015	0.0416	85.54	0.0382	0.0482	68.54	0.0485	0.0492
5	61.18	0.0410	0.0723	81.63	0.0032	0.0397	85.91	0.0377	0.0472	68.41	0.0507	0.0490
10	61.14	0.0404	0.0736	81.91	0.0048	0.0399	86.59	0.0384	0.0471	68.11	0.0498	0.0483

D.5 Efficiency and Scalability Study

In this section, we conduct our experiments with DP criterion to examine the communication cost and scalability of FedFACT.

Efficiency. We evaluate the communication efficiency of FedFACT by monitoring its performance across varying numbers of communication rounds T . As illustrated in Figure 4, the post-processing method, built upon a fully trained pre-trained model, consistently achieves convergence in fewer than 10 communication rounds, underscoring its high efficiency. The in-processing method likewise converges in under 40 iterations; given that it requires training the federated model from scratch, this performance is comparable to the convergence speed of FedAvg, making it highly effective compared to existing federated learning algorithms.

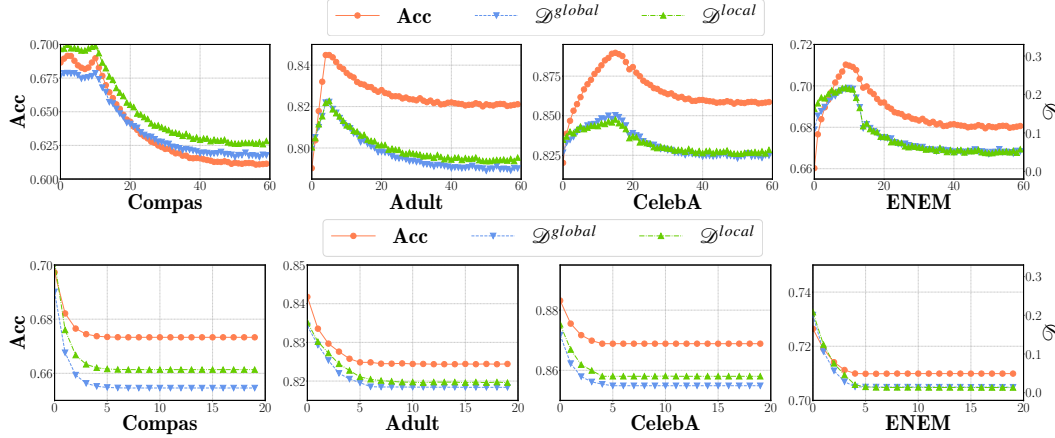


Figure 4: Communication Effectiveness Analysis. The convergence rates of both the in-processing (top row) and post-processing (bottom row) methods with respect to communication rounds on Compas, Adult, CelebA, and ENEM datasets.

Overall, whether employing the in-processing or post-processing method, all three performance metrics rapidly converge to stable values across each of the four datasets, empirically confirming both the communication efficiency and the overall effectiveness of FedFACT.

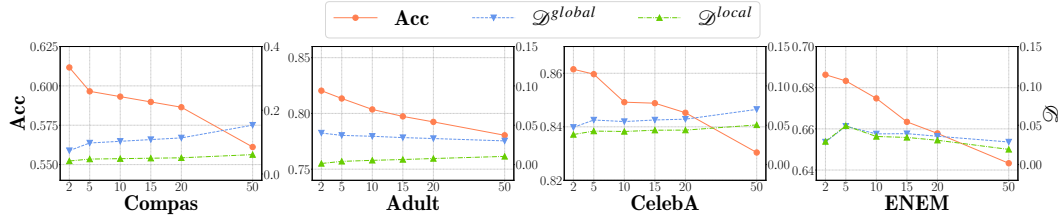
Scalability. We evaluate FedFACT’s performance as the number of clients varies from 2 to 50 on all four datasets, with heterogeneity parameter $\gamma = 5$ to ensure that each local client has adequate samples for assessing local fairness. The results, shown in Figure 5, indicate that on each dataset, there is an upward shift in the metric as the client count increases. Enforcing fairness constraints, especially via the in-processing method, sometimes necessitates a modest loss in accuracy, and the post-processing approach on the Compas dataset exhibits pronounced fairness fluctuations due to substantial generalization error when sample sizes are small. Aside from this, our method reliably bounds the model’s fairness, underscoring its robustness to variations in client population.

E Broader Impacts and Limitations

Broader Impacts. This paper addresses critical fairness issues in FL. By embedding fairness constraints at both the global and client levels, our framework delivers models that distribute accuracy more equitably, bolstering user confidence and mitigating bias amplification. The contributions of this research enhance user satisfaction and promote social equity. This fairness-aware approach extends readily to high-stakes classification tasks beyond FL: for instance, clinical decision support in hospital networks, vision-based detection systems, and financial fraud alerts. Integrating fairness into decentralized model training promotes privacy-preserving, equitable AI, helps satisfy emerging regulatory requirements, and encourages broader adoption of responsible machine learning across diverse application domains.

Limitations. The primary limitation of FedFACT is the fairness representation, which contains the linear disparities such as communly used DP, EOP and EO criteria, but it excludes some nonlinear formulations of fairness, e.g. Predictive Parity [20] and individual fairness [27]. Moreover, based on our generalization-error analysis, although the proposed method enables a controllable accuracy-fairness trade-off for a given fairness metric, it still requires a sufficiently large local sample size to accurately estimate local fairness (whereas global fairness demands only an adequate overall sample size). While our empirical results compare favorably against existing approaches, exploiting dataset

Behavior of The In-Processing Method under Varying Numbers of Clients.



Behavior of The Post-Processing Method under Varying Numbers of Clients.

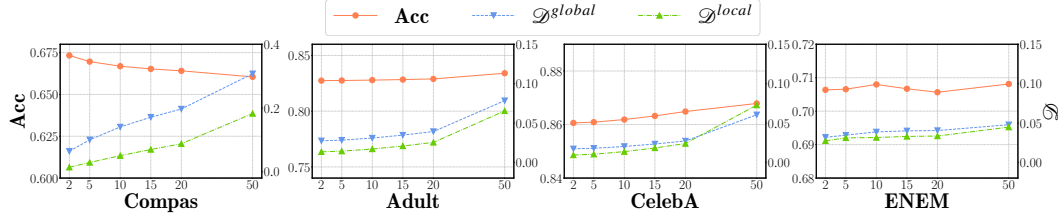


Figure 5: Scalability Analysis. The behavior of both the in-processing (top row) and post-processing (bottom row) methods as the number of clients increases from 2 to 50 across Compas, Adult, CelebA, and ENEM datasets.

1515 characteristics to optimize fairness may reduce the sample complexity needed for local fairness
 1516 optimization. Addressing these limitations remains an important avenue for future work.