
FedFACT: A Provable Framework for Controllable Group-Fairness Calibration in Federated Learning

Li Zhang¹, Zhongxuan Han¹, Xiaohua Feng¹, Jiaming Zhang¹, Yuyuan Li², Chaochao Chen^{1*}

¹Zhejiang University, ²Hangzhou Dianzi University

zhangliz180@gmail.com, {zxhan, fengxiaohua, 22321350}@zju.edu.cn
y2li@hdu.edu.cn, zjuccc@zju.edu.cn

Abstract

With the emerging application of Federated Learning (FL) in decision-making scenarios, it is imperative to regulate model fairness to prevent disparities across sensitive groups (e.g., female, male). Current research predominantly focuses on two concepts of group fairness within FL: *Global Fairness* (overall model disparity across all clients) and *Local Fairness* (the disparity within each client). However, the non-decomposable, non-differentiable nature of fairness criteria poses two fundamental, unresolved challenges for fair FL: (i) *Harmonizing global and local fairness, especially in multi-class setting*; (ii) *Enabling a controllable, optimal accuracy-fairness trade-off*. To tackle these challenges, we propose a novel controllable federated group-fairness calibration framework, named FedFACT. FedFACT identifies the Bayes-optimal classifiers under both global and local fairness constraints, yielding models with minimal performance decline while guaranteeing fairness. Building on the characterization of the optimal fair classifiers, we reformulate fair federated learning as a personalized cost-sensitive learning problem for in-processing and a bi-level optimization for post-processing. Theoretically, we provide convergence and generalization guarantees for FedFACT to approach the near-optimal accuracy under given fairness levels. Extensive experiments on multiple datasets across various data heterogeneity demonstrate that FedFACT consistently outperforms baselines in balancing accuracy and global-local fairness.

1 Introduction

Federated learning (FL) is a collaborative distributed machine learning paradigm that allows multiple clients to jointly train a shared model while preserving the privacy of their local data [55]. As FL is increasingly adopted in high-stakes domains—healthcare [80, 65, 62], finance [15, 11, 72], pattern recognition [52, 63, 94, 86, 19], and recommender systems [10, 82, 35, 73]—ensuring fairness is imperative to prevent discrimination against demographic groups based on sensitive attributes [32, 90, 88, 64], such as race, gender, age, etc. Although a rich literature addresses group fairness in centralized settings [2, 3, 41, 13], these methods depend on full access to the entire dataset and centralized processing, imposing excessive communication overhead and elevating privacy concerns when directly applied in the FL context.

To provide fairness guarantees for federated algorithms, recent works have concentrated on two group-fairness concepts in FL: *Global Fairness* and *Local Fairness* [33, 25, 95, 18, 31]. Global fairness aims to identify a model that provides similar treatment to protected groups across the entire data distribution. Local fairness concerns models that mitigate disparities and deliver unbiased predictions for sensitive groups within each client’s local data. Previous work [33] theoretically demonstrated that, under statistical heterogeneity across clients, global and local fairness can differ,

*Corresponding author

and both entail an inherent trade-off with predictive accuracy. As global and client-level biases can induce heterogeneous treatment disparities among sensitive groups, concurrently mitigating global and local disparities is vital for achieving group fairness in FL. For example, in constructing a federated prediction model for clinical decision-making within a hospital network [48], achieving global fairness substantially enhances performance for disadvantaged subgroups, while fairness at each hospital also carries heightened significance due to local deployment and legal requirements [21].

However, existing methods face certain challenges in controlling group fairness within FL: (i) *Harmonizing global and local fairness, especially in multi-class classification*. Divergent sensitive-group distributions from client heterogeneity separate global and local fairness, thereby imposing an intrinsic trade-off [33, 25]. Most fair FL approaches focus exclusively on either global or local fairness [31, 18, 93, 23, 85], thereby inevitably sacrificing the other objective and impeding the realization of both fairness criteria. Moreover, this research predominantly addresses fairness in the binary case, despite the ubiquity of multiclass tasks in practical FL scenarios. (ii) *Enabling a controllable, optimal accuracy-fairness trade-off with theoretical guarantee*. The non-decomposable, non-differentiable nature of group-fairness measures poses significant optimization challenges [56, 17]. Existing frameworks typically rely on surrogate fairness losses [85, 31, 93, 18, 54], yet the inevitable surrogate-fairness gap [83, 53] produces suboptimal performance and undermines convergence stability, thus hampering the controllability of the accuracy-fairness trade-off.

To address these challenges, we propose a novel **F**ederated group-**F**airness **C**alibra**T**ion framework, named FedFACT, comprising in-processing and post-processing approaches. Our framework is capable of ensuring controllable global and local fairness with minimal accuracy deterioration, underpinned by provable convergence and consistency guarantees. To harmonize global and local fairness, we seek to find the optimal classifier under both dual fairness constraints in the multi-class case. To this end, specific characterizations of the federated Bayes-optimal fair classifiers are established for both the in-processing and post-processing phases in FL. Building on the Bayes-optimal fair classifier’s structure, we develop efficient, privacy-preserving federated optimization strategies that realize a controllable and theoretically optimal fairness–accuracy trade-off. In detail, FedFACT reduces the in-processing task into a series of personalized cost-sensitive classification problems, and reformulates post-processing as a bi-level optimization that leverages the explicit form of the federated Bayes-optimal fair classifiers. We further derive theoretical convergence and generalization guarantees, demonstrating that our methods achieve near-optimal model performance while enforcing tunable global and local fairness constraints.

Our extensive experiments on multiple real-world datasets verify the efficiency and effectiveness of FedFACT, highlighting that FedFACT delivers a superior, controllable accuracy-fairness balance while maintaining competitive classification performance compared to state-of-the-art methods.

Our main contributions are summarized as follows:

- To the best of our knowledge, we are the first to propose a multi-class federated group-fairness calibration framework that approaches the Bayes-optimal fair classifiers, explicitly tailored to achieve a provably optimal and controllable balance between global fairness, local fairness, and accuracy.
- We further develop efficient algorithms to derive optimal classifiers under global-local fairness constraints at in-processing and post-processing stages with provable convergence and consistency guarantees. The in-processing fair classification is reduced to a sequence of personalized cost-sensitive learning problem, while the post-processing is formulated as a bi-level optimization, using the closed-form representation of Bayes-optimal fair classifiers.
- We conduct extensive experiments on multiple datasets with various data heterogeneity. The experimental results demonstrate that FedFACT outperforms existing methods in achieving superior balances among global fairness, local fairness, and accuracy. Experiments also show that FedFACT enables the flexible adjustment of the accuracy-fairness trade-off in FL.

2 Related Work

Group Fairness in Machine Learning. As summarized in previous work [56], group fairness is broadly defined as the absence of prejudice or favoritism toward a sensitive group based on their inherent characteristics. Common strategies for realizing group fairness in machine learning can be

classified into three categories: pre-, in-, and post-processing methods. **Pre-processing** [47, 42, 43] approaches aim to modify training data to eradicate underlying bias before model training. **In-processing** [45, 50, 2, 81, 83, 87, 92, 84] methods are developed to achieve fairness requirements by intervening during the training process. **Post-processing** [13, 20, 91, 78, 79, 37, 46] methods adjust the prediction results generated by a given model to adapt to fairness constraints after the training stage. However, because these methods require access to the full dataset, they are confined to mitigating disparities only at the local level.

Group Fairness in Federated Learning. Current methods primarily utilize in-processing and post-processing strategies to address global or local fairness issues. Concerning local fairness, prior work [12] highlights potential detrimental effects of FL on the group fairness of each clients, while [18] and [85] employ unified and personalized multi-objective optimization algorithms, respectively, to navigate the trade-off between local fairness and accuracy. Concerning global fairness, two main approaches are adaptive reweighting techniques [58, 31, 93, 1] and optimizing relaxed fairness objectives within FL [23, 75, 26], generally replacing the fairness metrics with surrogate functions.

Furthermore, previous work [33] offered a theoretical study elucidating the divergence between local and global fairness in FL, while revealing the intrinsic trade-off between these fairness objectives and accuracy. [25] formulates local and global fairness optimization into a linear program for minimal fairness cost, but it does not realize the Bayes-optimal balance between accuracy and fairness. [95] derives the Bayes-optimal classifier and decomposes the overall problem into client-specific optimizations, yet their approach applies only to binary classification in post-processing. Persistent challenges include inadequate accuracy-fairness flexibility and limited convergence guarantee in mitigating disparities within multi-class classification across various stages of FL training.

3 Preliminary

Notation. Denote by $\mathbf{1}_m$ the m -dimensional all-ones vector and by $\mathbf{1}_{m \times m}$ the $m \times m$ all-ones matrix. Write the probability simplex as $\Delta_m = \{q \in [0, 1]^m : \|q\|_1 = 1\}$, and let e_i be the i -th standard basis vector. Throughout this paper, bold lowercase letters denote vectors and bold uppercase letters denote matrices. For two equally sized matrices $\mathbf{A}$ and $\mathbf{B}$, their Frobenius inner product is $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i,j} a_{ij} b_{ij}$. For positive integer n , $[n] = \{1, \dots, n\}$.

Fairness in classification. Let (X, A, Y) be a random tuple, where $X \in \mathcal{X}$ for some feature space $\mathcal{X} \subseteq \mathbb{R}^d$, labels $Y \in \mathcal{Y} = [m]$, and the discrete sensitive attribute $A \in \mathcal{A}$. Given the randomized classifiers $h : \mathcal{X} \times \mathcal{A} \rightarrow \Delta_m$, the prediction $\hat{Y}$ is associated with the random outputs of h defined by $\mathbb{P}(\hat{Y} = y \mid \mathbf{x}) = h_y(\mathbf{x})$. In this work, we generally focus on three popular group-fairness criteria—Demographic Parity (DP) [27], Equalized Opportunity (EOP) [37] and Equalized Odds [37]—in multiclass classification tasks with multiple sensitive attributes, as defined in prior works [20, 78, 79].

Group fairness in FL. A federated system consists of numerous decentralized clients, so that we consider the population data distribution represented by a jointly random tuple (X, A, Y, K) with total N clients. The k -th client possesses a local data dataset $\mathcal{D}_k$, $k \in [N]$. Each sample in $\mathcal{D}_k$ is assumed to be drawn from local distribution, represented as $\{(x_{k,i}, a_{k,i}, y_{k,i})\}_{i=1}^{n_k}$, where n_k represents the number of samples for client k . The Bayes score function is commonly used to characterize the performance-optimal classifier under fairness constraints [91, 13, 78, 20], which possesses a natural extension in the federated setting: $\eta_y(x, a, c) := \mathbb{P}(Y = y \mid X = x, A = a, K = c)$. We are interested in both local and global fairness criteria in the FL context following [33, 25, 95]:

Definition 1. (Global Fairness) The disparity regarding sensitive groups aroused by the federated model in global data distribution $\mathbb{P}(X, A, Y)$ across all clients.

Definition 2. (Local Fairness) The disparity regarding sensitive groups aroused by the federated model when evaluated on each client’s data distribution $\mathbb{P}(X, A, Y \mid K)$.

Confusion matrices. Confusion matrices encapsulate the information required to evaluate diverse performance metrics and assess group fairness constraints in classification tasks [81, 60, 46]. The population confusion matrix is $\mathbf{C} \in [0, 1]^{m \times m}$, with elements defined for $i, j \in [m]$ as $\mathbf{C}_{i,j} = \mathbb{P}(Y = i, \hat{Y} = j)$. To capture both local and global fairness across multiple data distributions within

Table 1: Example of Confusion-Matrix-Based Group Fairness Constraints in Centralized & Federated Learning.

Fairness Criterion	Demographic parity (DP)	Equal Opportunity (EOP)
Group Constraints (Centralized)	$\left \mathbb{P}(\hat{Y} = y \mid A = a') - \mathbb{P}(\hat{Y} = y) \right $ $\forall a' \in \mathcal{A}, \forall y \in \mathcal{Y}$	$\left \mathbb{P}(\hat{Y} = y \mid A = a', Y = y) - \mathbb{P}(\hat{Y} = y \mid Y = y) \right $ $\forall a' \in \mathcal{A}, \forall y \in \mathcal{Y}$
Matrix Notations (Centralized)	$\left \sum_a \sum_i (\mathbb{I}[a = a'] - p_a) \mathbf{C}_{i,y}^a \right $ $\forall a' \in \mathcal{A}, \forall y \in \mathcal{Y}$	$\left \sum_a \left(\frac{p_{a',y}}{p_{a',y}} \mathbb{I}[a = a'] - \frac{1}{p_y} \right) \mathbf{C}_{y,y}^a \right $ $\forall a' \in \mathcal{A}, \forall y \in \mathcal{Y}$
Global Fairness (Federated)	$\left \sum_a \sum_k \sum_i (p_{k a'} \mathbb{I}[a = a'] - p_{a,k}) \mathbf{C}_{i,y}^{a,k} \right $ $\forall a' \in \mathcal{A}, \forall y \in \mathcal{Y}$	$\left \sum_a \sum_k \left(\frac{p_{a',k}}{p_{a',y}} \mathbb{I}[a = a'] - \frac{p_{a,k}}{p_{y,k}} \right) \mathbf{C}_{i,y}^{a,k} \right $ $\forall a' \in \mathcal{A}, \forall y \in \mathcal{Y}$
Local Fairness (Federated)	$\left \sum_a \sum_i (\mathbb{I}[a = a'] - p_{a k}) \mathbf{C}_{i,y}^{a,k} \right $ $\forall a' \in \mathcal{A}, \forall y \in \mathcal{Y}$	$\left \sum_a \sum_k \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} \right $ $\forall a' \in \mathcal{A}, \forall y \in \mathcal{Y}$

FL, we propose the **decentralized group-specific confusion matrices** $\mathbf{C}^{a,k}$, $a \in \mathcal{A}, k \in [N]$, with elements defined for $i, j \in [m]$ as $\mathbf{C}_{i,j}^{a,k}(h) := \mathbb{P}(Y = i, \hat{Y} = j \mid A = a, K = k)$.

Fairness and performance metrics. As presented in Table 1 (EO criterion and notation explanations see Appendix A), the global fairness constraints typically can be expressed by $|\mathcal{D}_{u_g}^g(h)| \leq \xi^g$, $u_g \in \mathcal{U}_g$, where $\mathcal{D}_{u_g}^g(h) = \sum_a \sum_k \langle \mathbf{D}_{u_g}^{a,k}, \mathbf{C}^{a,k}(h) \rangle$ represents the constraints required to achieve the global fairness criterion. Similarly, the local fairness constraints are $|\mathcal{D}_{u_k}^k(h)| \leq \xi^k$, $u_k \in \mathcal{U}_k, k \in [N]$, with $\mathcal{D}_{u_k}^k(h) = \sum_a \langle \mathbf{D}_{u_k}^{a,k}, \mathbf{C}^{a,k}(h) \rangle$. For performance metrics, we consider a risk metric expressed as a linear function of the population confusion matrix, i.e. $\mathcal{R}(h) = \langle \mathbf{R}, \mathbf{C}(h) \rangle = \sum_a \sum_k p_{a,k} \langle \mathbf{R}, \mathbf{C}^{a,k}(h) \rangle$. This formulation has been explored in multi-label and fair classification contexts [76, 81], and encompasses a variety of performance metrics, such as average recall and precision. In this paper, we primarily focus on standard classification error to set $\mathbf{R} = \mathbf{1}_{m \times m} - \mathbf{I}$.

4 Methodology

4.1 Federated Bayes-Optimal Fair Classifier

To investigate the optimal classifier with the group fairness guarantee within FL, we consider the situation that there is a unified fairness constraint at the global level, and each client has additional local fairness restrictions in response to personalized demands. Therefore, it is appropriate to consider a personalized federated model to minimize classification risk and ensure both local and global fairness. Denoting the set of classifiers as $\mathcal{H} = \{h : \mathcal{X} \rightarrow \Delta_m\}$, the federated Bayes-optimal fair classification problem can be formulated as

$$\min_{\mathbf{h} \in \mathcal{H}^N} \mathcal{R}(\mathbf{h}), \quad \text{s.t. } |\mathcal{D}^g(\mathbf{h})| \leq \xi^g, \quad |\mathcal{D}^k(\mathbf{h})| \leq \xi^k, \quad k \in [N], \quad (1)$$

where ξ^k, ξ^g are positive bounds, $\mathcal{D}^g(\mathbf{h}) := \{\mathcal{D}_{u_g}^g(\mathbf{h})\}_{u_g \in \mathcal{U}_g}$, $\mathcal{D}^k(\mathbf{h}) := \{\mathcal{D}_{u_k}^k(\mathbf{h})\}_{u_k \in \mathcal{U}_k}$, and the inequality applies element-wise. FL model $\mathbf{h} = (h_1, \dots, h_N)$ comprises N local classifiers.

Before delving into the optimal solution for Problem (1), we present a formal result on the structure of the federated Bayes-optimal fair classifier. Proposition 1 indicates that the Bayes-optimal classifier can be decomposed into local deterministic classifiers for all clients. This observation provides valuable insights for the subsequent algorithm design. The proof and the discussion on feasibility are given in Appendix B.1.

Proposition 1. *If (1) is feasible for any positive ξ^k and ξ^g , the client-wise classifier h_k^* in federated Bayes-optimal fair classifier $\mathbf{h}^* = (h_1^*, \dots, h_N^*)$ can be expressed as the linear combination of some deterministic classifiers $\{h'_{k,i}\}_{i=1}^{d_k}$, i.e., $h_k^*(x) = \sum_{i=1}^{d_k} \alpha_{k,i} h'_{k,i}(x)$, $\alpha_k \in \Delta_{d_k}$.*

4.2 In-processing Fair Federated Training via Cost Sensitive Learning

In this section, we aim to seek for the optimal solution of $\mathbf{h} = (h_1, \dots, h_N)$, where each local classifier is attribute-blind and parameterized by ϕ_k . A direct approach to solving (1) is to formulate an equivalent convex-concave saddle point problem in terms of its Lagrangian $\mathcal{L}(\mathbf{h}, \lambda, \mu)$:

$$\mathcal{R}(\mathbf{h}) + (\lambda^{(1)} - \lambda^{(2)})^\top \mathcal{D}^g(\mathbf{h}) - (\lambda^{(1)} + \lambda^{(2)})^\top \xi^g + \sum_{k=1}^N (\mu_k^{(1)} - \mu_k^{(2)})^\top \mathcal{D}^k(\mathbf{h}) - (\mu_k^{(1)} + \mu_k^{(2)})^\top \xi^k,$$

where $\lambda = \{\lambda^{(1)}, \lambda^{(2)}\}$ and $\mu = \{\mu_k^{(1)}, \mu_k^{(2)}\}_{k=1}^N$ are the dual parameters. Let $\Lambda := \{\lambda \in \mathbb{R}_{\geq 0}^{2|\mathcal{U}_g|} : \|\lambda\|_1 \leq B_d\}$ and $\mathcal{M} := \{\mu \in \mathbb{R}_{\geq 0}^{2\sum_k |\mathcal{U}_k|} : \|\mu\|_1 \leq B_d\}$. Since $\mathbf{h}$ is random classifier, by Sion's minimax theorem [66], the primal problem can be written as a saddle-point optimization

$$\max_{\lambda \in \Lambda, \mu \in \mathcal{M}} \min_{\mathbf{h} \in \mathcal{H}^N} \mathcal{L}(\mathbf{h}, \lambda, \mu) = \min_{\mathbf{h} \in \mathcal{H}^N} \max_{\lambda \in \Lambda, \mu \in \mathcal{M}} \mathcal{L}(\mathbf{h}, \lambda, \mu). \quad (2)$$

The boundedness of optimal λ^*, μ^* will be shown later. To derive the representation of optimal saddle point, we initially focus on the inner minimization optimization task, namely $\min_{\mathbf{h} \in \mathcal{H}} \mathcal{L}(\mathbf{h}, \lambda, \mu)$.

Proposition 2. *Given non-negative λ and μ , then an optimal solution $\mathbf{h}^* = (h_1^*, \dots, h_N^*)$ to the inner problem $\min_{\mathbf{h} \in \mathcal{H}} \mathcal{L}(\mathbf{h}, \lambda, \mu)$ is realized by local deterministic classifiers $h_k^*(x)$, $k \in [N]$ satisfying*

$$h_k^*(x) = e_y, y \in \arg \max_{j \in [m]} \left(\sum_{a \in \mathcal{A}} \mathbb{P}(A = a | x, k) [\mathbf{M}^{\lambda, \mu}(a, k)]^\top \eta(x, a, k) \right)_j, \quad (3)$$

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

The proof of Proposition 2 is given in Appendix B.2. Notice that the optimal solution in (3) remains computationally intractable, because the point-wise distributions $\mathbb{P}(A = a | x, k)$ and the Bayes-optimal classifier η are unknown. We reformulate the task of solving $\mathbf{h}^*$ within a cost-sensitive learning framework, by designing sample-wise calibrated training losses for each $h_k^*(x)$ that can yield an equivalent objective.

Proposition 3. *Let the personalized cost-sensitive loss for client k be defined by*

$$\ell_k(y, \mathbf{s}(x), a) = - \sum_{i=1}^m \bar{\mathbf{M}}_{y,i}^{\lambda, \mu}(a, k) \log \frac{\exp([\mathbf{s}(x)]_i)}{\sum_{j=1}^m \exp([\mathbf{s}(x)]_j)}, \quad (4)$$

where $\mathbf{s} : \mathcal{X} \rightarrow \mathbb{R}^m$ is the scoring function, and $\bar{\mathbf{M}}^{\lambda, \mu}(a, k) = \mathbf{M}^{\lambda, \mu}(a, k) + \kappa \mathbf{1}_{m \times m}$ with κ chosen to ensure all matrix entries are strictly positive. Denoting the optimal scoring function to minimize ℓ_k over the local data distribution as $\mathbf{s}_k^*(x)$, then the loss ℓ_k is calibrated for the inner problem $\min_{\mathbf{h} \in \mathcal{H}} \mathcal{L}(\mathbf{h}, \lambda, \mu)$, i.e., $h_k^*(x) = e_y, y \in \arg \max_{j \in [m]} [\mathbf{s}_k^*(x)]_j$ is equivalent to that in (3).

In practice, $\mathbf{s}$ is parameterized by ϕ_k for $k \in [N]$, formulating the personalized optimization objective $L_k(f_{\phi_k}) = \sum_{i=1}^{n_k} \ell_k(y_{k,i}, \mathbf{s}(x_{k,i}; \phi_k), a_{k,i})$ in the FL setting. Appendix B.3 provides the proof of Proposition 3 and further presents that the loss ℓ_k in (4) is also calibrated for the unified federated Bayes-optimal fair classifier. Inspired by this property, we propose an efficient in-processing algorithm for group-fair classification within FL, as detailed in Algorithm 1.

At each iteration t , the personalized classifier h_k^t is obtained by ensembling the unified model θ^t with the local model ϕ_k^t . The ensemble weight w_k^t and its update rule balance the contributions of the unified and local models. The following theorem establishes the personalized regret bound w.r.t. the best model parameter, and further demonstrates that our algorithm achieves an ϵ -approximate stochastic saddle point.

Theorem 4. *Under mild assumptions, there exist constants B_k, B_L , such that the following cumulative regret upper bound is guaranteed for the ensemble personalized models:*

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}[L_k^t(f_{k,ens}^t) - L_k^t(f_{\phi_k^*}^t)] \leq \frac{\|\phi_k^*\|^2}{\eta RT} + \eta B_k + \frac{\log(2)}{\eta_w T} + \eta_w B_L. \quad (5)$$

Furthermore, suppose that personalized models achieve a ρ^t -approximate optimal response at iteration 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 $\bar{\rho}^T = \sum_{t=1}^T \rho^t / T$, then the sequences of model and bounded dual parameters comprise an approximate mixed Nash equilibrium:

$$\max_{\lambda^*, \mu^*} \frac{1}{T} \sum_{t=1}^T \hat{\mathcal{L}}(\mathbf{h}^t, \lambda^*, \mu^*) - \inf_{\mathbf{h}^*} \frac{1}{T} \sum_{t=1}^T \hat{\mathcal{L}}(\mathbf{h}^*, \lambda^t, \mu^t) \leq \epsilon^T := \bar{\rho}^T + 16B_d^2 \sqrt{\frac{1}{T}}. \quad (6)$$

The complete Theorem 4 with its proof is provided in the Appendix B.4. The regret bound yields a convergence rate of $\mathcal{O}(1/\sqrt{T})$ by appropriately choosing the learning rate, reflecting the stability of the proposed algorithm. Moreover, as ρ^t decreases with t , the algorithm will gradually converge to the optimal equilibrium. For instance, if $\rho^t \propto C/\sqrt{t}$, ϵ will also exhibit an $\mathcal{O}(1/\sqrt{T})$ convergence rate. The **generalization error** between the optimal solutions of the empirical dual and primal problems under finite samples is given in **Appendix B.5**.

Algorithm 1: FedFACT (In-processing)

Input : Datasets $\{x_{k,i}, y_{k,i}, a_{k,i}\}_{i=1}^{N_k}$ from client $k, k \in \mathcal{K}$; Communication round T ; Local round R ; Initial parameters $\lambda^0, \mu^0, \theta^0, \phi_k^0, w_k^0, k \in \mathcal{K}$; Learning rate $\{\eta, \eta_d, \eta_w\}_{t=1}^T$;
for $t = 0, 1, \dots, T$ **do**
 Each client $k \in \mathcal{K}$ **in parallel do**:
 Ensemble unified and local model: $h_k^t(x) = e_y, y \in \arg \max_{j \in [m]} [f_{k,ens}^t(x)]_j$, where
 $f_{k,ens}^t(x) = w_k^t f_{\theta^t}(x) + (1 - w_k^t) f_{\phi_k^t}(x)$ and $f_{\phi}(x) := \text{softmax}(\mathbf{s}(x; \phi))$;
 Update calibration matrix $\bar{\mathbf{M}}^{\lambda^t, \mu_k^t} = \widehat{\mathbf{M}}^{\lambda^t, \mu_k^t} + \kappa \mathbf{1}_{m \times m}$;
 Update the weight $w_k^{t+1} = \frac{1}{1 + \mathcal{W}_k^t(w_k^t)}, \mathcal{W}_k^t(w_k^t) = \frac{1 - w_k^t}{w_k^t} \exp(-\eta_w [L_k(f_{\phi_k^t}) - L_k(f_{\theta^t})])$;
 Calculate update of global dual parameter $\Delta \lambda_k^{t+1}$, and update local dual parameter μ_k^{t+1} ,
 $\Delta \lambda_{k,u_g}^{(i),t+1} = (3 - 2i) \sum_{a \in \mathcal{A}} \langle \widehat{\mathbf{D}}_{u_g}^{a,k}, \widehat{\mathbf{C}}^{a,k}(h_k^t) \rangle - \xi^g, i \in [1, 2], u_g \in \mathcal{U}_g$,
 $\mu_{k,u_k}^{(i),t+1} = \Pi_{\mathcal{M}}[(3 - 2i) \sum_{a \in \mathcal{A}} \langle \widehat{\mathbf{D}}_{u_k}^{a,k}, \widehat{\mathbf{C}}^{a,k}(h_k^t) \rangle - \xi^k], i \in [1, 2], u_k \in \mathcal{U}_k$;
 Perform R local-batch update of θ^t and ϕ_k^t , guided by loss L_k with $\bar{\mathbf{M}}^{\lambda^t, \mu_k^t}$,
 $\theta_k^{t,r+1} = \theta_k^{t,r} - \eta_k \nabla L_k(\theta^{t,r}; \mathcal{B}_k^{t,r}), \phi_k^{t,r+1} = \phi_k^{t,r} - \eta_k \nabla L_k(\phi^{t,r}; \mathcal{B}_k^{t,r}), r = 0, \dots, R - 1$;
 Send last update $\Delta \theta_k^{t+1} = \theta_k^{t,R} - \theta^t$ to the server;
 Server do:
 Server aggregates $\{\Delta \theta_k^{t+1}\}$: $\theta^{t+1} = \theta^t + \sum_{k=1}^N p_k \Delta \theta_k^{t+1}$;
 Update global dual parameter: $\lambda^{t+1} = \Pi_{\Lambda}(\lambda^t + \eta_d \sum_{k=1}^N \Delta \lambda_k^{t+1})$;
 Send $\theta^{t+1}, \lambda^{t+1}$ to clients;
end
Return Personalized classifier $\bar{\mathbf{h}} = (\bar{h}_1, \dots, \bar{h}_N)$, where $\bar{h}_k := \sum_{t=1}^T h_k^t / T$;

4.3 Label-Free Federated Post-Fairness Calibration based on Plug-In Approach

This section introduces a post-hoc fairness approach that calibrates the classification probabilities of a pre-trained federated model. We formulate an closed-form representation of the federated Bayes optimal fair classifier under standard assumptions, and then derive the primal problem into bi-level optimization through the plug-in approach. To begin, we introduce the following assumption.

Assumption 1. (η -continuity). For each client k , denoting $\mathcal{P}_{a,k}^X := \mathbb{P}(X \mid A = a, K = k)$, let the put forward distribution $\tau_{a,k} := \eta_a^* \mathcal{P}_{a,k}^X, a \in \mathcal{A}$ be absolutely continuous with respect to the Lebesgue measure restricted to Δ_m .

Assumption 1, which can be met by adding minor random noises to $\tau_{a,k}$, is commonly used in the literature on post-processing fairness [13, 81, 20, 16]. Next, we derive a more explicit characterization of federated Bayes-optimal fair classifier.

Theorem 5. Under Assumption 1, suppose that the primal problem (1) is feasible for any $\xi^g, \xi^k > 0$ and all non-zero columns of each $\mathbf{D}_u^{a,k}$ are distinct, the attribute-aware personalized classifier $\{h_k^*\}_{k \in [N]}$ is Bayes-optimal under local and global fairness constraints, if $h_k^*(x, a) = e_y, y \in \arg \max_{j \in [m]} \left([\mathbf{M}^{\lambda^*, \mu^*}(a, k)]^\top \eta(x, a, k) \right)_j$, where the dual parameters are determined from $(\lambda^*, \mu^*) \in \arg \min_{\lambda, \mu \geq 0} H(\lambda, \mu)$,

$$H(\lambda, \mu) = \sum_{k \in [N]} \sum_{a \in \mathcal{A}} p_{a,k} \mathbb{E}_{X \sim \mathcal{P}_{a,k}^X} \left[\max_{y \in [m]} \left([\mathbf{M}^{\lambda, \mu}(a, k)]^\top \eta(X, a, k) \right)_y \right] + \xi^g \|\lambda\|_1 + \sum_{k \in [N]} \xi^k \|\mu_k\|_1. \quad (7)$$

The optimal dual parameters (λ^*, μ^*) are bounded, and the optimality of λ and μ respectively guarantee global fairness and local fairness.

Proof of Theorem 5 is given in Appendix B.6. Since the dual parameter μ only related to local fairness constraints, each clients can finish update of this parameter without global aggregation. Therefore,

the federated Bayes-optimal classification problem can be reformulated into a bi-level optimization:

$$\min_{\lambda \in \Lambda} \left\{ \hat{F}(\lambda) = \sum_{k=1}^N \hat{p}_k \hat{F}_k(\lambda) \right\}, \quad \hat{F}_k(\lambda) := \min_{\mu_k \in \mathcal{M}_k} \hat{H}_k(\lambda, \mu_k), \quad (8)$$

where $\hat{H}_k(\lambda, \mu_k) := \frac{1}{n_k} \sum_{i=1}^{n_k} \max_{y \in [m]} ([\mathbf{M}^{\lambda, \mu}(a_i, k)]^\top \eta(x_i, a_i, k))_y + \xi^g \|\lambda\|_1 + \frac{\xi^k}{\hat{p}_k} \|\mu_k\|_1$ is the plug-in estimation of (7). Considering that the non-smoothness of the optimization objective may lead to convergence issues in federated optimization [89], we replace the maximum operation in $\hat{H}_k(\lambda, \mu_k)$ with soft-max weight function $\sigma_\beta(x) = \sum_{i=1}^m \frac{\exp(x_i/\beta)}{\sum_{j=1}^m \exp(x_j/\beta)} x_i$, which reduces to the hard-maximum if temperature $\beta \rightarrow 0$. Denoting the relaxed local objective as $\hat{H}'_k(\lambda, \mu_k)$, we propose Algorithm 2 to solve the federated Bayes-optimal fair classifier.

Algorithm 2: FedFACT (Post-processing)

Input : Datasets $\mathcal{D}_k = \{x_{k,i}, y_{k,i}, a_{k,i}\}_{i=1}^{N_k}$ from client $k \in [N]$; Communication round T ;

Local round R ; Initial parameters $\lambda^0, \mu^0, k \in [N]$; Learning rate η_d ; Pre-trained $\hat{\eta}$

for $t = 0, 1, \dots, T$ **do**

Each client $k \in [N]$ **in parallel do:**

 Perform R local-batch update of $\mu_k^t, i = 1, 2$ with $\hat{H}'_k(\lambda, \mu_k)$:

$$\mu_k^{t,r+1} = \Pi_{\mathcal{M}_k}(\mu_k^{t,r} - \eta_d \nabla_{\mu} \hat{H}'_k(\lambda, \mu_k)), r = 0, \dots, R-1.$$

 Set $\mu_k^{t+1} = \mu_k^{t,R}$, and calculate update of λ^t : $\Delta \lambda_k^{t+1} = \nabla_{\lambda} \hat{H}'_k(\lambda^t, \mu_k^{t+1})$.

Server do:

 Server aggregates $\{\Delta \lambda_k^{t+1}\}$: $\lambda^{t+1} = \Pi_{\Lambda}(\lambda^t - \eta_d \sum_{k=1}^N \hat{p}_k \Delta \lambda_k^{t+1})$; Send λ^{t+1} to clients;

end

Return Classifiers $\{h_1, \dots, h_N\}, h_k(x, a) := \arg \max_{j \in [m]} ([\widehat{\mathbf{M}}^{\lambda^*, \mu^*}(a, k)]^\top \hat{\eta}(x, a, k))_j$;

Proposition 6. *The bi-level objectives $\hat{H}'_k(\lambda, \mu_k), k \in [N]$ are convex and L -smooth.*

Proof of Proposition 6 is given in Appendix B.7. Existing research in FL [74] shows that the L -smoothness of the local objective suffice for Algorithm 2 to achieve an $\mathcal{O}(1/\sqrt{T})$ convergence rate. Moreover, owing to the equivalence of nested and joint minimization under convexity [66, 36], the corresponding bi-level optimization can approach the optimal solution of the empirical primal problem. Consequently, the remaining error arises from the **generalization risk** induced by finite sampling, which is explored in **Appendix B.8**.

4.4 Discussion

In- versus post-processing. In- and post-processing interventions play complementary roles: the former removes bias in representations during training (incurring higher computational cost), and the latter adjusts fairness on model outputs with low overhead, unable to debias learned representations. Both of them support adaptable fairness calibration in resource-limited, heterogeneous FL. Note that combining the in-processing and post-processing methods is theoretically unjustifiable, as the in-processing classifier is not designed to approximate the Bayes score function.

Efficiency & Privacy. Each iteration of our in- and post-processing methods requires only a single client-server interaction and is supported by convergence guarantees that demonstrate our algorithms' efficiency. This will be empirically validated in our experiments; FedFACT is also privacy-friendly. In-processing requires sharing λ alongside the unified model θ , while post-processing involves sharing only λ . These exchanges conform to standard FL [55] and preserve data confidentiality. Furthermore, differential privacy [28, 30] or encryption schemes [7] can be applied to further reinforce privacy.

5 Experiments

To comprehensively assess the proposed FedFACT framework, we conduct extensive experiments on four publicly available real-world datasets to answer the following Research Questions (RQ): **RQ1:**

Table 2: Overall Experimental Results.

Partition	Dataset	Compas			Adult			CelebA			ENEM		
		Acc	$\mathcal{G}^{global}$	$\mathcal{G}^{local}$	Acc	$\mathcal{G}^{global}$	$\mathcal{G}^{local}$	Acc	$\mathcal{G}^{global}$	$\mathcal{G}^{local}$	Acc	$\mathcal{G}^{global}$	$\mathcal{G}^{local}$
$\gamma = 0.5$	FedAvg	69.73	0.2766	0.3590	84.52	0.1765	0.2310	89.14	0.1435	0.1308	67.56	0.2620	0.2462
	FairFed	59.39	0.1008	0.1022	80.73	0.0983	0.1434	81.85	0.0704	0.1058	60.99	0.1165	0.1733
	FedFB	58.09	0.0879	0.0983	81.85	0.0751	0.1165	85.32	0.1188	0.0949	64.35	0.0814	0.1326
	FCFL	56.53	0.0646	0.0614	81.91	0.0845	0.1455	83.83	0.0704	0.1090	61.07	0.1260	0.1189
	praFFed	59.93	0.0824	0.0968	80.96	0.0591	0.0763	85.45	0.0731	0.0862	62.60	0.0736	0.0806
	Cost	64.51	0.0585	0.0941	81.09	0.0262	0.0590	85.60	0.0314	0.0577	65.79	0.0487	0.0674
	FedFACT _g (In)	61.19	0.0344	0.0761	82.05	0.0015	0.0408	86.49	0.0205	0.0544	63.79	0.0493	0.0568
	FedFACT _l (In)	61.81	0.0600	0.0636	82.44	0.0140	0.0508	85.90	0.0461	0.0312	63.93	0.0434	0.0487
	FedFACT _{g&l} (In)	61.17	0.0407	0.0732	82.04	0.0014*	0.0401	86.15	0.0382	0.0473	62.51	0.0366	0.0413
	FedFACT _g (Post)	67.27	0.0128*	0.0660	82.83*	0.0173	0.0276	<u>87.25</u>	0.0089*	0.0253	66.15	<u>0.0175</u>	0.0181*
	FedFACT _l (Post)	67.49*	0.0315	0.0552*	82.79	0.0154	0.0267*	87.36*	0.0127	0.0163*	66.54*	0.0197	0.0240
	FedFACT _{g&l} (Post)	<u>67.33</u>	<u>0.0139</u>	<u>0.0641</u>	82.74	0.0134	<u>0.0274</u>	87.06	<u>0.0093</u>	<u>0.0172</u>	<u>66.52</u>	0.0162*	<u>0.0184</u>
Hetero	FedAvg	69.12	0.2513	0.4044	85.34	0.1596	0.2358	89.85	0.1360	0.1742	67.45	0.2037	0.3068
	FairFed	61.87	0.1825	0.2448	81.85	0.1074	0.1283	82.50	0.0672	0.1415	59.81	0.0949	0.1624
	FedFB	60.16	0.1284	0.1332	81.14	0.0949	0.1005	86.00	0.1121	0.1284	63.68	0.0780	0.2039
	FCFL	59.96	0.1498	0.1507	79.10	0.0528	0.0596	84.50	0.0657	0.1458	61.05	0.1134	0.1425
	praFFed	60.42	0.0902	0.1062	80.12	0.0523	0.0606	85.13	0.0592	0.1152	60.84	0.0932	0.1246
	Cost	63.01	0.0773	0.1044	81.04	0.0286	0.0567	86.28	0.0495	0.0761	62.11	0.0315	0.0730
	FedFACT _g (In)	60.33	0.0665	0.0841	82.09*	0.0122	0.0962	86.18	0.0188	0.0731	62.05	0.0380	0.0577
	FedFACT _l (In)	60.22	0.0730	0.0753	81.19	0.0208	<u>0.0250</u>	86.58	0.0424	0.0426	61.96	0.0471	0.0473
	FedFACT _{g&l} (In)	61.44	0.0676	0.0789	81.19	0.0055	<u>0.0239*</u>	86.38	0.0355	0.0634	61.48	0.0322	0.0364
	FedFACT _g (Post)	64.36	0.0398*	<u>0.0699</u>	81.09	0.0047*	0.0257	<u>87.95</u>	0.0094	0.0344	<u>65.32</u>	0.0199*	0.0343
	FedFACT _l (Post)	<u>64.38</u>	0.0479	0.0740	81.27	<u>0.0049</u>	0.0306	88.05*	0.0112	0.0217*	65.68*	0.0246	0.0120*
	FedFACT _{g&l} (Post)	64.41*	<u>0.0408</u>	0.0680*	81.31	0.0053	0.0293	87.75	0.0088*	<u>0.0247</u>	65.19	<u>0.0201</u>	0.0131

* The best results are marked with *. The second-best results are underlined.

* All results are the average of five repeated experiments.

* We use **FedAvg** as the baseline for optimal accuracy, without comparing its accuracy-fairness trade-off.

Does FedFACT outperform the existing methods in effectively achieving a global-local accuracy-fairness balance? **RQ2:** Is FedFACT capable of adjusting the trade-off between accuracy and global-local fairness (sensitivity analysis)? **RQ3:** How do important hyper-parameters influence the performance of FedFACT? **RQ4:** How about the communication efficiency and scalability of FedFACT?

5.1 Datasets and Experimental Settings

Due to space limitations, the detailed information in this section is provided in **Appendix C**.

Datasets. Experiments are conducted on four real-world datasets: **Compas**[22], **Adult** [4], **CelebA** [96], and **ENEM** [40], which are well established for assessing fairness in FL [31, 12, 24, 95].

Baselines. For binary classification, experiments are conducted on all four datasets. We compare our method with traditional federated baselines **FedAvg** [55] and five state-of-the-art methods tailored for addressing global and local fairness within FL, namely **FairFed** [31], **FedFB** [93], **FCFL** [18], **praFFed** [85] and the method in [25], denoted as **Cost** in our experiments. The reason for we did not include the experiments with [95] is explained in Appendix C.2. For multi-group or multi-class classification, the experiments are implemented on CelebA and ENEM datasets due to label limitations.

Data distribution. To model the statistical heterogeneity in the FL context, we investigate two data partitioning strategies: (i) *Dirichlet partition*: we control the distribution of the sensitive attribute at each client using a Dirichlet distribution $Dir(\gamma)$ as proposed by [31]. A smaller γ indicates greater heterogeneity across clients. (ii) *Heterogeneous split*: Inspired by [33], we propose a partitioning method that introduces heterogeneous correlations between the sensitive attribute A and label Y . The correlation between A and Y for each client is controlled by a parameter randomly sampled from $[0, 1]$, as detailed in Appendix C.3.

Evaluation. (i) *Firstly*, we partition each dataset into a 60% training set and the remaining 40% for test set. (ii) *Secondly*, the number of clients is set to 2 in Compas, and 5 in other datasets to ensure sufficient samples for local fairness estimation. (iii) *Thirdly*, we evaluate the FL model with Accuracy (Acc), global fairness metric ($\mathcal{G}^{global}$), and maximal local fairness metric among clients ($\mathcal{G}^{local}$), with smaller values of fairness metrics indicating a fairer FL model.

5.2 Overall Comparison (RQ1)

We perform extensive experiments comparing FedFACT against existing fair FL baselines under varying statistical heterogeneity. We set $\xi^g = \xi^k = 0.01$ for FedFACT. The subscript g and l

Table 3: Accuracy-Fairness Balance (Sensitivity Analysis).

Dataset (ξ^g, ξ^l)	Compas (In-)			Adult (In-)			Compas (Post-)			Adult (Post-)		
	Acc	$\mathcal{G}^{global}$	$\mathcal{G}^{local}$	Acc	$\mathcal{G}^{global}$	$\mathcal{G}^{local}$	Acc	$\mathcal{G}^{global}$	$\mathcal{G}^{local}$	Acc	$\mathcal{G}^{global}$	$\mathcal{G}^{local}$
(0.00,0.00)	61.17	0.0407	0.0732	82.04	0.0014	0.0401	67.33	0.0139	0.0641	82.74	0.0134	0.0274
(0.02,0.00)	61.39	0.0548	0.0848	82.37	0.0028	0.0458	67.49	0.0315	0.0552	82.75	0.0154	0.0255
(0.04,0.00)	61.81	0.0600	0.0836	82.44	0.0140	0.0508	67.92	0.0557	0.0692	82.83	0.0173	0.0276
(0.00,0.02)	61.23	0.0418	0.0732	82.04	0.0018	0.0409	67.40	0.0134	0.0658	82.76	0.0139	0.0278
(0.02,0.02)	61.50	0.0569	0.0895	82.41	0.0056	0.0450	67.93	0.0558	0.0624	82.77	0.0166	0.0262
(0.04,0.02)	61.63	0.0665	0.0933	82.52	0.0080	0.0479	67.95	0.0623	0.0598	82.81	0.0150	0.0256
(0.00,0.02)	61.23	0.0418	0.0732	82.04	0.0014	0.0408	67.27	0.0128	0.0660	82.79	0.0154	0.0267
(0.02,0.04)	61.66	0.0556	0.0919	82.57	0.0089	0.0442	67.95	0.0536	0.0644	82.81	0.0174	0.0283
(0.04,0.04)	62.39	0.0720	0.1105	82.66	0.0223	0.0449	68.03	0.0645	0.0598	82.81	0.0185	0.0278

represent the presence of global and local fairness constraints, respectively. It is essential to note that there is an inherent trade-off between accuracy and global-local fairness.

Binary classification results. We compare FedFACT with benchmarks tailored to binary classification in terms of the binary DP and EOP criteria on the four datasets. The results of DP are presented in Table 2. We also report the Pareto frontier in **Appendix D.1** to evaluate the ability of FedFACT to strike a accuracy-fairness balance, along with the Pareto results of binary EOP criterion. Overall, when compared to existing SOTA methods, FedFACT demonstrates superior performance in achieving a balanced trade-off between accuracy and fairness.

Multi-Class results. We illustrate how FedFACT performs on multi-class prediction using CelebA and ENEM with DP and EOP constraints. As there are no established methods addressing multi-class fairness in federated learning, we conduct comparisons only between FedAvg and our proposed in- and post-processing approaches, as shown in Figure 1. More experimental details are in Appendix D.2.

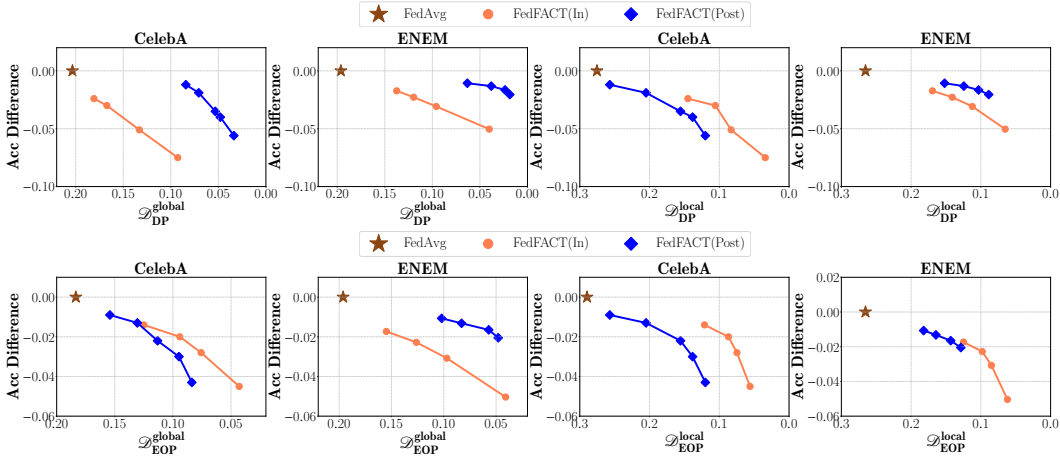


Figure 1: Multi-Class Fair Classification Results. The top line depict global and local multiclass Demographic Parity (DP) results, while the bottom line show global and local multiclass Equal Opportunity (EOP) outcomes.

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. 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.

5.3 Flexibility of Adjusting Accuracy-Fairness Trade-Off (RQ2)

To investigate the capability of FedFACT in adjusting accuracy-fairness trade-off, we examine the Acc, $\mathcal{G}^{global}$ and $\mathcal{G}^{local}$ under different fairness relaxation of (ξ^g, ξ^l) with $\gamma = 0.5$ on Adult and Compas in Table 3. Here we set the local fairness levels ξ^k for each client to the same value, denoted as ξ^l . More experimental results are presented in Appendix D.3.

Sensitivity Analysis. Table 3 shows that, for a fixed global constraint ξ^g , reducing ξ^l diminishes both accuracy and local fairness—implying that stricter local fairness comes at the cost of overall performance. Conversely, by keeping ξ^l constant, one can modulate global fairness via adjustments to ξ^g . Note that the difference between the constraints and the fairness metrics arises due to the unavoidable generalization error with finite samples. In general, these findings substantiate our claim that FedFACT enables flexible control over the accuracy-fairness trade-off in FL.

5.4 Hyper-parameter Experiments (RQ3)

There is no tunable hyper-parameter in our proposed method except for **the number of deterministic classifiers** utilized to construct the weight classifiers. We gradually raise the number of classifiers forming the weighted classifier, starting with the most recent one and extending to the previous 10 classifiers. The detailed experimental results are provided in **Appendix D.4**.

5.5 Efficiency and Scalability Study (RQ4)

In **Appendix D.5**, we undertake extensive experiments to empirically demonstrate the communication efficiency and scalability to client number of the proposed method FedFACT.

6 Conclusion

This paper introduces a novel controllable **F**ederated group-**F**airness **C**alibra**T**ion framework called FedFACT, to ensure both group and local fairness within FL. FedFACT is proposed to learn the federated Bayes-optimal fair classifier in both in- and post-processing stages, which achieves a theoretically minimal accuracy loss with both fairness constraints. We developed efficient algorithms—with convergence and consistency guarantees—that reduce fair classification to personalized cost-sensitive learning for in-processing and bi-level optimization for post-processing. Extensive experiments on four publicly available real-world datasets demonstrate that FedFACT outperforms SOTA methods, exhibiting a remarkable ability to harmonious balance between accuracy and global-local fairness.

Reproducibility Statement

Details for the experimental setting are provided in the beginning of Section 5 and Appendix C, and the code can be found at <https://github.com/liizhang/FedFACT>.

Acknowledgments

This work was supported in part by the Hangzhou Key Scientific Research Plan (No. 2024SZD1A28), and National Natural Science Foundation of China (No. 62402148).

References

- [1] Annie Abay, Yi Zhou, Nathalie Baracaldo, Shashank Rajamoni, Ebube Chuba, and Heiko Ludwig. Mitigating bias in federated learning, 2020.
- [2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR, 2018.
- [3] Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, Peter Michalak, Shahab Asoodeh, and Flavio Calmon. Beyond adult and compas: Fair multi-class prediction via information projection. *Advances in Neural Information Processing Systems*, 35:38747–38760, 2022.
- [4] Arthur Asuncion, David Newman, et al. Uci machine learning repository, 2007.
- [5] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

- [6] Dimitri P. Bertsekas. Stochastic optimization problems with nondifferentiable cost functionals. *Journal of Optimization Theory and Applications*, 12:218–231, 1973.
- [7] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS '17*, page 1175–1191, New York, NY, USA, 2017. Association for Computing Machinery.
- [8] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.
- [9] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [10] Robin Burke. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093*, 2017.
- [11] David Byrd and Antigoni Polychroniadou. Differentially private secure multi-party computation for federated learning in financial applications. In *Proceedings of the first ACM international conference on AI in finance*, pages 1–9, 2020.
- [12] Hongyan Chang and Reza Shokri. Bias propagation in federated learning. *arXiv preprint arXiv:2309.02160*, 2023.
- [13] Wenlong Chen, Yegor Klochkov, and Yang Liu. Post-hoc bias scoring is optimal for fair classification. In *The Twelfth International Conference on Learning Representations*, 2024.
- [14] Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. A fair classifier using kernel density estimation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA, 2020. Curran Associates Inc.
- [15] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.
- [16] Evgenii Chzhenn, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. *Advances in Neural Information Processing Systems*, 32, 2019.
- [17] Andrew Cotter, Heinrich Jiang, Maya Gupta, Serena Wang, Taman Narayan, Seungil You, and Karthik Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 20(172):1–59, 2019.
- [18] Sen Cui, Weishen Pan, Jian Liang, Changshui Zhang, and Fei Wang. Addressing algorithmic disparity and performance inconsistency in federated learning. *Advances in Neural Information Processing Systems*, 34:26091–26102, 2021.
- [19] Ittai Dayan, Holger R Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Z Abidin, Andrew Liu, Anthony Beardsworth Costa, Bradford J Wood, Chien-Sung Tsai, et al. Federated learning for predicting clinical outcomes in patients with covid-19. *Nature medicine*, 27(10):1735–1743, 2021.
- [20] Christophe Denis, Romuald Elie, Mohamed Hebiri, and François Hu. Fairness guarantees in multi-class classification with demographic parity. *Journal of Machine Learning Research*, 25(130):1–46, 2024.
- [21] Department of Health and Human Services, Centers for Medicare & Medicaid Services, Office of the Secretary. Nondiscrimination in health programs and activities. Federal Register, vol. 89, no. 87, pp. 37522–37703, May 6, 2024. Document No. 2024-08711.
- [22] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

- [23] Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. Fairness-aware agnostic federated learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 181–189. SIAM, 2021.
- [24] Yuying Duan, Yijun Tian, Nitesh Chawla, and Michael Lemmon. Post-fair federated learning: Achieving group and community fairness in federated learning via post-processing. *arXiv preprint arXiv:2405.17782*, 2024.
- [25] Yuying Duan, Gelei Xu, Yiyu Shi, and Michael Lemmon. The cost of local and global fairness in federated learning. *arXiv preprint arXiv:2503.22762*, 2025.
- [26] Gerry Windiartha Mohamad Dunda and Shenghui Song. Fairness-aware federated minimax optimization with convergence guarantee, 2024.
- [27] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.
- [28] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.
- [29] Ahmad-Reza Ehyaei, Golnoosh Farnadi, and Samira Samadi. Wasserstein distributionally robust optimization through the lens of structural causal models and individual fairness. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 42430–42467. Curran Associates, Inc., 2024.
- [30] Ahmed El Ouadrhiri and Ahmed Abdelhadi. Differential privacy for deep and federated learning: A survey. *IEEE access*, 10:22359–22380, 2022.
- [31] Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A Salman Avestimehr. Fairfed: Enabling group fairness in federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 7494–7502, 2023.
- [32] Laura Gustafson, Chloe Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng-Yang Fu, Melissa Hall, and Candace Ross. Facet: Fairness in computer vision evaluation benchmark. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20313–20325, 2023.
- [33] Faisal Hamman and Sanghamitra Dutta. Demystifying local & global fairness trade-offs in federated learning using partial information decomposition. In *The Twelfth International Conference on Learning Representations*, 2024.
- [34] Xiaotian Han, Jianfeng Chi, Yu Chen, Qifan Wang, Han Zhao, Na Zou, and Xia Hu. FFB: A fair fairness benchmark for in-processing group fairness methods. In *The Twelfth International Conference on Learning Representations*, 2024.
- [35] Zhongxuan Han, Chaochao Chen, Xiaolin Zheng, Li Zhang, and Yuyuan Li. Hypergraph convolutional network for user-oriented fairness in recommender systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 903–913, 2024.
- [36] Filip Hanzely and Peter Richtárik. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020.
- [37] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.
- [38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [39] Xiaolin Hu, Shaojie Li, and Yong Liu. Generalization bounds for federated learning: Fast rates, unparticipating clients and unbounded losses. In *The Eleventh International Conference on Learning Representations*, 2023.

- [40] COMITÊ DE ÉTICA INEP. Instituto nacional de estudos e pesquisas educacionais anísio teixeira. *Boletim de Serviço Eletrônico em*, 30:04, 2018.
- [41] Nikola Jovanović, Mislav Balunovic, Dimitar Iliev Dimitrov, and Martin Vechev. Fare: Provably fair representation learning with practical certificates. In *International Conference on Machine Learning*, pages 15401–15420. PMLR, 2023.
- [42] Nikola Jovanović, Mislav Balunovic, Dimitar Iliev Dimitrov, and Martin Vechev. FARE: Provably fair representation learning with practical certificates. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15401–15420. PMLR, 23–29 Jul 2023.
- [43] Jian Kang, Jingrui He, Ross Maciejewski, and Hanghang Tong. Inform: Individual fairness on graph mining. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD ’20, page 379–389, New York, NY, USA, 2020. Association for Computing Machinery.
- [44] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR, 10–15 Jul 2018.
- [45] Dongha Kim, Kunwoong Kim, Insung Kong, Ilsang Ohn, and Yongdai Kim. Learning fair representation with a parametric integral probability metric. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11074–11101. PMLR, 17–23 Jul 2022.
- [46] Joon Sik Kim, Jiahao Chen, and Ameet Talwalkar. FACT: A diagnostic for group fairness trade-offs. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5264–5274. PMLR, 13–18 Jul 2020.
- [47] Peizhao Li and Hongfu Liu. Achieving fairness at no utility cost via data reweighing with influence. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12917–12930. PMLR, 17–23 Jul 2022.
- [48] Siqi Li, Qiming Wu, Xin Li, Di Miao, Chuan Hong, Wenjun Gu, Yuqing Shang, Yohei Okada, Michael Hao Chen, Mengying Yan, et al. Fairfml: Fair federated machine learning with a case study on reducing gender disparities in cardiac arrest outcome prediction. *arXiv preprint arXiv:2410.17269*, 2024.
- [49] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pages 6357–6368. PMLR, 2021.
- [50] Tianlin Li, Qing Guo, Aishan Liu, Mengnan Du, Zhiming Li, and Yang Liu. FAIRER: Fairness as decision rationale alignment. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19471–19489. PMLR, 23–29 Jul 2023.
- [51] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.
- [52] Yuxi Liu, Guibo Luo, and Yuesheng Zhu. Fedfms: Exploring federated foundation models for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 283–293. Springer, 2024.

- [53] Michael Lohaus, Michael Perrot, and Ulrike Von Luxburg. Too relaxed to be fair. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6360–6369. PMLR, 13–18 Jul 2020.
- [54] Disha Makhija, Xing Han, Joydeep Ghosh, and Yejin Kim. Achieving fairness across local and global models in federated learning. *arXiv preprint arXiv:2406.17102*, 2024.
- [55] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.
- [56] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.
- [57] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.
- [58] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International conference on machine learning*, pages 4615–4625. PMLR, 2019.
- [59] Harikrishna Narasimhan and Aditya Krishna Menon. Training over-parameterized models with non-decomposable objectives. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS ’21, Red Hook, NY, USA, 2021. Curran Associates Inc.
- [60] Harikrishna Narasimhan, Harish G. Ramaswamy, Shiv Kumar Tavker, Drona Khurana, Praneeth Netrapalli, and Shivani Agarwal. Consistent multiclass algorithms for complex metrics and constraints. *Journal of Machine Learning Research*, 25(367):1–81, 2024.
- [61] Lawrence Narici and Edward Beckenstein. *Topological vector spaces*. Chapman and Hall/CRC, 2010.
- [62] Dinh C Nguyen, Quoc-Viet Pham, Pubudu N Pathirana, Ming Ding, Aruna Seneviratne, Zihuai Lin, Octavia Dobre, and Won-Joo Hwang. Federated learning for smart healthcare: A survey. *ACM Computing Surveys (Csur)*, 55(3):1–37, 2022.
- [63] Sarthak Pati, Ujjwal Baid, Brandon Edwards, Micah Sheller, Shih-Han Wang, G Anthony Reina, Patrick Foley, Alexey Gruzdev, Deepthi Karkada, Christos Davatzikos, et al. Federated learning enables big data for rare cancer boundary detection. *Nature communications*, 13(1):7346, 2022.
- [64] Yangyang Qu, Michele Panariello, Massimiliano Todisco, and Nicholas Evans. Reference-free adversarial sex obfuscation in speech. *arXiv preprint arXiv:2508.02295*, 2025.
- [65] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020.
- [66] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer Verlag, Heidelberg, Berlin, New York, 2009.
- [67] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fairbatch: Batch selection for model fairness. In *International Conference on Learning Representations*, 2021.
- [68] Milad Sefidgaran, Romain Chor, Abdellatif Zaidi, and Yijun Wan. Lessons from generalization error analysis of federated learning: You may communicate less often! In *Forty-first International Conference on Machine Learning*, 2024.
- [69] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.
- [70] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

- [71] Jacob Steinhardt and Percy Liang. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *International conference on machine learning*, pages 1593–1601. PMLR, 2014.
- [72] Qianyi Sun, Zheyong Qiu, Hong Ye, and Zhiyao Wan. Multinational corporation location plan under multiple factors. In *Journal of Physics: Conference Series*, volume 1168, page 032012. IOP Publishing, 2019.
- [73] Zehua Sun, Yonghui Xu, Yong Liu, Wei He, Lanju Kong, Fangzhao Wu, Yali Jiang, and Lizhen Cui. A survey on federated recommendation systems. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):6–20, 2024.
- [74] Davoud Ataee Tarzanagh, Mingchen Li, Christos Thrampoulidis, and Samet Oymak. FedNest: Federated bilevel, minimax, and compositional optimization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 21146–21179. PMLR, 17–23 Jul 2022.
- [75] Ganghua Wang, Ali Payani, Myungjin Lee, and Ramana Kompella. Mitigating group bias in federated learning: Beyond local fairness. *arXiv preprint arXiv:2305.09931*, 2023.
- [76] Xiaoyan Wang, Ran Li, Bowei Yan, and Oluwasanmi Koyejo. Consistent classification with generalized metrics, 2019.
- [77] Robert C. Williamson, Elodie Vernet, and Mark D. Reid. Composite multiclass losses. *Journal of Machine Learning Research*, 17(222):1–52, 2016.
- [78] Ruicheng Xian, Lang Yin, and Han Zhao. Fair and optimal classification via post-processing. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 37977–38012. PMLR, 23–29 Jul 2023.
- [79] Ruicheng Xian and Han Zhao. A unified post-processing framework for group fairness in classification, 2024.
- [80] Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated learning for healthcare informatics. *Journal of healthcare informatics research*, 5:1–19, 2021.
- [81] Forest Yang, Mouhamadou Cisse, and Sanmi Koyejo. Fairness with overlapping groups; a probabilistic perspective. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4067–4078. Curran Associates, Inc., 2020.
- [82] Liu Yang, Ben Tan, Vincent W Zheng, Kai Chen, and Qiang Yang. Federated recommendation systems. In *Federated Learning: Privacy and Incentive*, pages 225–239. Springer, 2020.
- [83] Wei Yao, Zhanke Zhou, Zhicong Li, Bo Han, and Yong Liu. Understanding fairness surrogate functions in algorithmic fairness. *Transactions on Machine Learning Research*, 2024.
- [84] Mehdi Yazdani-Jahromi, Ali Khodabandeh Yalabadi, AmirArsalan Rajabi, Aida Tayebi, Ivan Garibay, and Ozlem Garibay. Fair bilevel neural network (fairbinn): On balancing fairness and accuracy via stackelberg equilibrium. *Advances in Neural Information Processing Systems*, 37:105780–105818, 2024.
- [85] Rongguang Ye, Wei-Bin Kou, and Ming Tang. Praffl: A preference-aware scheme in fair federated learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD ’25, page 1797–1808, New York, NY, USA, 2025. Association for Computing Machinery.
- [86] Zhenyu Yu and Chee Seng Chan. Yuan: Yielding unblemished aesthetics through a unified network for visual imperfections removal in generated images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9716–9724, 2025.

- [87] Zhenyu Yu, Mohd Yamani Idna Idris, and Pei Wang. Physics-constrained symbolic regression from imagery. In *2nd AI for Math Workshop@ ICML 2025*, 2025.
- [88] Zhenyu Yu, Mohd Yamani Idna Idris, Pei Wang, Yuelong Xia, and Yong Xiang. Forgetme: Benchmarking the selective forgetting capabilities of generative models. *Engineering Applications of Artificial Intelligence*, 161:112087, 2025.
- [89] Honglin Yuan, Manzil Zaheer, and Sashank Reddi. Federated composite optimization. In *International Conference on Machine Learning*, pages 12253–12266. PMLR, 2021.
- [90] Boya Zeng, Yida Yin, and Zhuang Liu. Understanding bias in large-scale visual datasets. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 61839–61871. Curran Associates, Inc., 2024.
- [91] Xianli Zeng, Guang Cheng, and Edgar Dobriban. Bayes-optimal fair classification with linear disparity constraints via pre-, in-, and post-processing. *arXiv preprint arXiv:2402.02817*, 2024.
- [92] Yi Zeng, Xuelin Yang, Li Chen, Cristian Ferrer, Ming Jin, Michael Jordan, and Ruoxi Jia. Fairness-aware meta-learning via nash bargaining. *Advances in Neural Information Processing Systems*, 37:83235–83267, 2024.
- [93] Yuchen Zeng, Hongxu Chen, and Kangwook Lee. Improving fairness via federated learning. *arXiv preprint arXiv:2110.15545*, 2021.
- [94] J. Zhang, W. Zhang, C. Tan, X. Li, and Q. Sun. Yolo-ppa based efficient traffic sign detection for cruise control in autonomous driving. *arXiv preprint arXiv:2409.03320*, 2024.
- [95] Li Zhang, Chaochao Chen, Zhongxuan Han, Qiyong Zhong, and Xiaolin Zheng. Logofair: Post-processing for local and global fairness in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 22470–22478, 2025.
- [96] Yuanhan Zhang, ZhenFei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 70–85. Springer, 2020.
- [97] Zhaowei Zhu, Yuanshun Yao, Jiankai Sun, Hang Li, and Yang Liu. Weak proxies are sufficient and preferable for fairness with missing sensitive attributes. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 43258–43288. PMLR, 23–29 Jul 2023.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: In Section 1

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: In Appendix E

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: The theorems and proofs presented in both the Section 4 and the Appendix B are comprehensive and complete.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: Code is provided in the supplemental material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code is provided in the supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: In Section 5 and Appendix C

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: In Section 5 and Appendix C

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: In Appendix C.1

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: In Section E

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This question is not applicable as the paper does not release any data or models with a high risk of misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: In Section 5

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: Provided in the supplemental material.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: This question is not applicable as the paper does not involve crowdsourcing experiments or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[NA\]](#)

Justification: This question is not applicable as the paper does not release any data or models with a high risk of misuse.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

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 [81, 14, 44] to reduce the number of constraints, while our confusion-matrix representation is also applicable to the group-wise definitions of these fairness metrics [78, 79].

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 [81, 60] 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 EOP 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 $\text{FPR}_y(a) = \mathbb{P}(\hat{Y} = y | Y \neq y, A = a)$, $\text{FPR}_y = \mathbb{P}(\hat{Y} = y | Y \neq y)$.

It shows that

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

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 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 for the y -th row with all other elements set to 0 for $\mathbf{D}_{a',y}^{a,1} \in \mathbb{R}^{m \times m}$.

A.2 Group Fairness notations in FL

As noted in the main text, fairness at the level of each client's dataset (*local fairness*) differs from fairness across the aggregate dataset of all clients (*global fairness*). Local fairness is defined with respect to each client's individual data distribution $\mathbb{P}(X, Y, A \mid K)$, whereas global fairness is defined over the overall (aggregate) distribution $\mathbb{P}(X, Y, A)$. Motivated by approaches that employ group-specific confusion matrices for fairness [81, 60], we propose the **decentralized group-specific confusion matrices** $\mathbf{C}^{a,k}$, $a \in \mathcal{A}, k \in [N]$ to capture both global and local fairness across multiple data distributions within FL, with elements defined for $i, j \in [m]$ as $\mathbf{C}_{i,j}^{a,k}(h) := \mathbb{P}(Y = i, \hat{Y} = j \mid A = a, K = k)$.

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

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

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

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

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 $\mathbb{P}(K = k \mid A = a') \mathbb{I}[a = a'] - \mathbb{P}(A = a, K = k)$ with all other elements set to 0.

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

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

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

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

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 $\mathbb{I}[a = a'] - \mathbb{P}(A = a \mid K = k)$ with all other elements set to 0.

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

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

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

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

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 all other elements set to 0.

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

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

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 $\mathbb{P}(\hat{Y} = y \mid Y = y, K = k) = \sum_{a \in \mathcal{A}} \frac{p_{a,k}}{p_{y,k}} \mathbb{P}(\hat{Y} = y, Y = i \mid A = a, K = k)$. Hence, we have

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

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 all other elements set to 0.

Example 6. For EOP criterion, the global EOP fairness metric is defined as

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

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

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

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 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 for the y -th row with all other elements set to 0 for $\mathbf{D}_{a',y}^{a,k,1} \in \mathbb{R}^{m \times m}$.

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

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

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 $\mathbb{P}(\hat{Y} = y \mid Y \neq y, K = k) = \sum_{a \in \mathcal{A}} \frac{p_{a,k}}{\sum_{y_j \neq y} p_{y_j,k}} \mathbb{P}(\hat{Y} = y, Y = i \mid A = a, K = k)$. Hence, we have

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

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

A.3 Other Fairness Notations

Fairness notions formulated as ratios metrics can be converted into linear constraints under certain conditions, and our framework is well suited to enforce these fairness constraints in federated learning environments. Specifically, the ratios metric or constraints, which are formulated as $\left| \frac{\sum_a \langle \mathbf{D}^a, \mathbf{C}^a(h) \rangle}{\sum_a \langle \mathbf{G}^a, \mathbf{C}^a(h) \rangle} \right| \leq \xi$ with constant matrix $\mathbf{D}^a$, $\mathbf{G}^a$ and group-specific confusion matrix $\mathbf{C}^a(h)$ depend on classifier h , can certainly be transformed into multiple linear constraints if the sign of $\sum_a \langle \mathbf{G}^a, \mathbf{C}^a(h) \rangle$ is unchanged for any h . When the denominator's sign is uncertain, the feasible domain of $\mathbf{C}^a(h)$ is non-convex, precluding its expression via linear constraints. In fact, since each entry of $\mathbf{C}^a(h)$ lies in $[0,1]$, whenever the entries of $\mathbf{G}^a$ are sign-consistent, the corresponding ratio constraint admits a linear-constraint representation. For example, the Calibration within Groups (CG) which was proposed in [5] and further explored in [46], is a fairness metric in binary classification and can be formulated as $\frac{FN^a}{FN^a + TN^a} = v_0$; $\frac{TP^a}{TP^a + FP^a} = v_1$, where TP^a, FP^a, FN^a, TN^a are derived from binary group-specific confusion matrix $\mathbf{C}^a$ and $0 \leq v_0 < v_1 \leq 1$ and have no implicit dependence on any entries of the fairness-confusion tensor. Because this fairness criterion appears as a ratio metric and every element of corresponding $\mathbf{G}^a$ is non-negative, it admits a linear-constraint representation and can be realized in our proposed distributed framework. Moreover, the ratio metrics presented in [3] also can be formulated into multiple linear constraints based on the above analysis.

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 [81], Proposition 10 of [60]).

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 $\mathcal{D}^g, \mathcal{D}^k, k \in [N]$ 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 [9] 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 [61], 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 [9] 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 $\xi^g = 0, \xi^k = 0, k \in [N]$, 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 [60]. 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 [81].

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

problem (1),

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

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

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 non-negative dual parameter λ and μ , which can be formulated as

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

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

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 manipulations of H to reveal its clear relationship with the personalized classifier $\mathbf{h} = (h_1, \dots, h_N)$. Denote the condition distribution of X given sensitive attribute $A = a$ on client $K = k$ as $\mathcal{P}_{a,k}^X$, i.e., $\mathcal{P}_{a,k}^X := \mathbb{P}(X | A = a, K = k)$, we have

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

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

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

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

B.3 Proof of Proposition 3 and Further Exploration

In this section, we prove that the representation in (4) is calibrated for both unified and personalized 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*

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

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

This lemma is commonly used in the design of multiclass loss functions [77, 57, 59].

B.3.1 Proof of Proposition 3

Proof. We aim to prove that for any fixed $x \in \mathcal{X}, k \in [N]$, the optimal personalized scoring 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 distribution $\mathbb{P}(X, A, Y \mid K = k)$ recovers the personalized federated Bayes-optimal classifier $h_k^*(x)$ in Proposition 2.

It is equivalent to show that, for any x :

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

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

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

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

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

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, given fixed x, k , an optimal personalized classifier $\mathbf{s}_k^*(x)$ minimizing the cost-sensitive loss point-wise satisfies

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

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

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

Hence,

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

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

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

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

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

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

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

Consider the calibrated loss function in (4),

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

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

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

Hence,

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

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

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

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

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

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

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

Assumption 2. The local loss function $L_1^t, \dots, L_N^t$ are convex, β -smooth and bounded by B_L to 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:

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

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

Assumption 2 and 3 are standard in the convergence analysis of federated model [51, 49, 74]. Now we present Theorem 7 and Theorem 8, which together constitute an extended form of Theorem 4.

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

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

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 personalized regret upper bound:

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

Theorem 8. Suppose that personalized models achieve a ρ_t -approximate optimal response at iteration 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 model and bounded dual parameters comprise an approximate mixed Nash equilibrium:

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

B.4.1 Proof of Theorem 7

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

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

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

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 as

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

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

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

By convexity, we have

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

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

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

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

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

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

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

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

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

From (10), we know that

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

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

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

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

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

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

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

Here, the update can be viewed as exponentiated gradient descent on the normalized weight vector $\mathbf{w}_k^t = (w_{k,1}^t, w_{k,2}^t) \in \Delta_2$, and $w_{k,i}^t \propto \exp(-\eta_w z_{t,i}^k)$, $i = 1, 2$, where $z_{t,1}^k = L_k^t(f(x; \theta^t))$, $z_{t,2}^k = L_k^t(f(x; \phi_k^t))$. A well-known regret bound in online learning [70, 71] shows that, for any $\mathbf{u} = (u_1, u_2) \in \Delta_2$,

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

By the convexity of L_k^t , we have $\sum_{t=1}^T L_k^t(f_{k,ens}^t) \leq \sum_{t=1}^T w_k^t L_k^t(f(x; \theta^t)) + (1 - w_k^t) L_k^t(f(x; \phi_k^t))$. Plugging in $u_1 = 1, u_2 = 0$, it presents that

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

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

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

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

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

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

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

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

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

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

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

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

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

B.4.2 Proof of Theorem 8

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

Lemma B.2. [70] 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 $(\lambda^{(t)})$, and Π_Λ projects its argument onto Λ w.r.t. the Euclidean norm. Then:

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

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

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

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

where $\widehat{\mathbf{M}}^{\lambda, \mu}(a, k) := \mathbf{I} - \frac{1}{\widehat{p}_{a, k}} \left[\sum_{u_g \in \mathcal{U}_g} (\lambda_{u_g}^{(1)} - \lambda_{u_g}^{(2)}) \widehat{\mathbf{D}}_{u_g}^{a, k} - \sum_{u_k \in \mathcal{U}_k} (\mu_{k, u_k}^{(1)} - \mu_{k, u_k}^{(2)}) \widehat{\mathbf{D}}_{u_k}^{a, k} \right]$. It is clear that the inner problem is linear to classifiers in the empirical case.

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

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

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

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

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

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

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

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 $\bar{\lambda} := \frac{1}{T} \sum_{t=1}^T \lambda^t, \bar{\mu} := \frac{1}{T} \sum_{t=1}^T \mu^t$ denote the point-wise average of dual parameters. Therefore, due to the linearity of the empirical Lagrange function to classifiers and dual parameters, (25) can be formulated as

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

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

B.5 Generalization Error For In-processing Algorithm

We begin by introducing some notations and simplifications, which are commonly employed in generalization analyses of FL [39, 68]. Without loss of generalization, let $n = n_1 = \dots = n_k$ present the sample number in local datasets. For any class $\mathcal{H} = \{h : \mathcal{X} \rightarrow [m]\}$, denote $\mathcal{H}_y = \{\mathbb{I}\{h(x) = y\} : h \in \mathcal{H}\}$ and the maximal Vapnik-Chervonenkis dimension [69], $VC(\mathcal{H}) := \max_{y \in [m]} VC(\mathcal{H}_y)$.

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 of empirical Lagrangian $\widehat{\mathcal{L}}(\mathbf{h}, \lambda, \mu)$, and an optimal solution $\mathbf{h}^* \in \mathcal{H}$ satisfies both global and local fairness constraints, denoting $\nu(n, \mathcal{H}, \delta) = 2\sqrt{\frac{2VC(\mathcal{H}) \log(n+1)}{n}} + \sqrt{\frac{2 \log(m^2 N / \delta)}{n}}$, $B_g = \max_{a \in \mathcal{A}, k \in [N]} \|\mathbf{D}_{u_g}^{a, k}\|_1, \Omega_n^g = \max_{a \in \mathcal{A}, k \in [N]} \|\mathbf{D}_{u_g}^{a, k} - \widehat{\mathbf{D}}_{u_g}^{a, k}\|_\infty, \Omega_n^p := \sum_{k=1}^N |p_k - \widehat{p}_k|$, and $B_k = \max_{a \in \mathcal{A}} \|\mathbf{D}_{u_k}^{a, k}\|_1, \Omega_n^k = \max_{a \in \mathcal{A}} \|\mathbf{D}_{u_k}^{a, k} - \widehat{\mathbf{D}}_{u_k}^{a, k}\|_\infty, k \in [N]$, then with probability at least $1 - \delta$,*

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

The proof of Theorem 9 relies on the following lemma.

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. 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 at least $1 - \delta$, for $\forall i, j \in [m]$,*

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

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

Proof of Lemma B.3. Let $\ell_{i,j}(x, y; h) = \mathbb{I}(y = i \wedge h(x) = j)$. Then we have $\mathbf{C}_{i,j}(h) = \mathbb{E}[\ell_{i,j}(x, y; h)]$ 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 respect to cost sensitive binary classification [8], with probability at least $1 - \delta$,

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

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

B.5.1 Proof of Theorem 9

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

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

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

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

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

The last inequality is by the Holder's inequality. Let $\ell_{i,j}^{a',k}(x, y, a; h) = \mathbb{I}(y = i \wedge h(x) = j \wedge a = a')$. 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 union bound in Lemma B.3, we have that with probability at least $1 - \delta$,

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

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

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

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

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

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

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 constraints. With (27), we obtain

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

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

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

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

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

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

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

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

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

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

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

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

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

By (27) and (28),

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

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

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

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

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

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

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

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

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

This completes the proof. $\square$

B.6 Proof of Theorem 5

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

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

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

Lemma B.4. [6] The subdifferential of the function $F(x) = \mathbb{E}\{f(x, \omega)\}$ at a point x is given by

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

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

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

Lemma B.5. [66] Let $f_1, \dots, f_m : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ be convex functions. Define $f(x) = \max\{f_1(x), \dots, f_m(x)\}$, $\forall x \in \mathbb{R}^n$. For $x_0 \in \bigcap_{i=1}^m \text{dom} f_i$, define $I(x_0) = \{i \mid f_i(x_0) = f(x_0)\}$. Then $\partial f(x_0) = \text{conv} \bigcup_{i \in I(x_0)} \partial f_i(x_0)$.

Lemma B.6. [66] Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a convex continuous function. We consider the minimizer x^* of the function f over the set B . Then for x^* to be locally optimal it is necessary that

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

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 exists a subgradient $\xi \in \partial f(x^*)$, such that for all $k \in [d]$ we have $\xi_k \geq 0$ and $\forall k \in \mathcal{K}, \xi_k = 0$.

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

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

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

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

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

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

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

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

Before exploring the optimal solution of outer optimization, we first prove that the optimal dual parameter $\lambda^* \in \mathbb{R}_{\geq 0}^{2|\mathcal{U}_g|}$, $\mu^* \in \mathbb{R}_{\geq 0}^{2\sum_{k=1}^N |\mathcal{U}_k|}$ 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 $\forall \xi^g, \xi^k > 0$, 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 ξ^g, ξ^k . It is clear that for $\forall \xi^g, \xi^k > 0$, 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 $\xi^g, \xi^k > 0$. 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 $\xi^g, \xi^k > 0$. Let $\xi^g \rightarrow 0, \xi^k \rightarrow 0$, 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 a lower bound

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

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 abuse 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 $|\mathcal{D}_{u_k}^k(\mathbf{h}^{\lambda^*, \mu^*})| \leq \xi^k, k \in [N]$.

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 $\mathcal{D}^k(\mathbf{h}^{\lambda^*, \mu^*})$. 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. $\square$

B.7 Proof of Proposition 6

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

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

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

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

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 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 $\|x\|_1 \leq R$,

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

Hence, its spectral norm is bounded,

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

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

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

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

B.8 Generalization Error For Post-Processing Algorithm

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

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 of the empirical plug-in estimation of (7), denoting $\rho(n, \delta) = \sqrt{\frac{8|\mathcal{A}|m^2 \log(n+1)}{n}} + \sqrt{\frac{2 \log(m^2 |\mathcal{A}| N / \delta)}{n}}$, $B_g = \max_{a \in \mathcal{A}, k \in [N]} \|\mathbf{D}_{u_g}^{a, k}\|_1$, $\hat{B}_g = \max_{a \in \mathcal{A}, k \in [N]} \|\widehat{\mathbf{D}}_{u_g}^{a, k}\|_1$, $\Omega_n^g = \max_{a \in \mathcal{A}, k \in [N]} \|\mathbf{D}_{u_g}^{a, k} -$

$\widehat{\mathbf{D}}_{u_g}^{a,k} \|_\infty$, and $B_k = \max_{a \in \mathcal{A}} \|\mathbf{D}_{u_k}^{a,k}\|_1$, $\widehat{B}_k = \max_{a \in \mathcal{A}} \|\widehat{\mathbf{D}}_{u_k}^{a,k}\|_1$, $\Omega_n^k = \max_{a \in \mathcal{A}} \|\mathbf{D}_{u_k}^{a,k} - \widehat{\mathbf{D}}_{u_k}^{a,k}\|_\infty$, $k \in [N]$.

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

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

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

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

The proof of Theorem 10 needs the following lemma.

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

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

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

B.8.1 Proof of Theorem 10

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

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 ,

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

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

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 Lemma B.3, we have that with probability at least $1 - \delta$,

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

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

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

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

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

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

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

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

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

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 $b_t \in [0, 1]$. According to Carathéodory's theorem, the subgradient interval can still be represented by two points. According to our assumption, the plug-in estimator $\widehat{\eta}$ still meet the continuity assumption and we exclude singular λ', μ' . Therefore, we know that

$$\mathbb{P} \left(\sum_{i=1}^{n_{a,k}} \mathbb{I} \left(\exists t \neq y, \widehat{f}_t(x_i, a, k) \geq \widehat{f}_i(x_i, a, k), \forall i \in [m]; \widehat{f}_t(x_i, a, k) = \widehat{f}_y(x_i, a, k) \right) \leq 1 \right) = 1 \quad (56)$$

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

$$\frac{\partial}{\partial \lambda_{u_g}^{(1)}} \hat{H}(\lambda, \mu) \leq \sum_{k \in [N]} \sum_{a \in \mathcal{A}} \frac{1}{n_{a,k}} \sum_{i=1}^{n_{a,k}} \left[-[\hat{\eta}(x_i, a, k)]^\top \hat{\mathbf{D}}_{u_g}^{a,k} \hat{h}_k(x_i, a) \right] + \xi^g \quad (57)$$

$$+ \sum_{k \in [N]} \sum_{a \in \mathcal{A}} \frac{1}{n_{a,k}} \hat{B}_g \quad (58)$$

$$\leq -\hat{\mathcal{D}}_{u_g}^g(\hat{\mathbf{h}}) + \xi^g + \frac{|\mathcal{A}|N\hat{B}_g}{n_{\min}} \quad (59)$$

On the other hand,

$$\frac{\partial}{\partial \lambda_{u_g}^{(1)}} \hat{H}(\lambda, \mu) \geq \sum_{k \in [N]} \sum_{a \in \mathcal{A}} \frac{1}{n_{a,k}} \sum_{i=1}^{n_{a,k}} \left[-[\hat{\eta}(x_i, a, k)]^\top \hat{\mathbf{D}}_{u_g}^{a,k} \hat{h}_k(x_i, a) \right] + \xi^g \quad (60)$$

$$- \sum_{k \in [N]} \sum_{a \in \mathcal{A}} \frac{1}{n_{a,k}} \hat{B}_g \quad (61)$$

$$= -\hat{\mathcal{D}}_{u_g}^g(\hat{\mathbf{h}}) + \xi^g - \frac{|\mathcal{A}|N\hat{B}_g}{n_{\min}} \quad (62)$$

Hence, we obtain that

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

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

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

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 $0 \in \frac{\partial}{\partial \lambda_{u_g}^{(1)}} \hat{H}(\hat{\lambda}^*, \mu), 0 \in \frac{\partial}{\partial \lambda_{u_g}^{(2)}} \hat{H}(\hat{\lambda}^*, \mu)$. Thus,

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

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

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 $\hat{\lambda}_{u_g}^{(1)} = 0, \hat{\lambda}_{u_g}^{(2)} > 0$, as discussed in the proof of Theorem 5, it turns out that

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

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

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

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

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

□

C Additional Datasets and Experimental Setting

C.1 Datasets and Experimental Details

C.1.1 Datasets

- The **Compas** dataset [22] 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 [40] 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 [96] 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 [38] on CelebA as the classification model.

The determination of sensitive attributes and labels on three datasets has been verified significant in previous research [3, 34].

C.1.2 Baselines

We compare the performance of FedFACT with traditional **FedAvg** [55] and five SOTA methods tailored for calibrating global and local fairness in FL, namely **FairFed** [31], **FedFB** [93], **FCFL** [18], **praFFL** [85], and the method in [25], 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 [67] 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.
- [25] 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. 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 [85] and the global and local fairness constraints in Cost [25], 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 [95]

LogoFair [95] 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 [95]. 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.

C.3 Heterogeneous Split of Client Distribution

We propose a partitioning method that introduces heterogeneous correlations between the sensitive attribute A and label Y , thereby further elucidating the trade-off between global fairness and local fairness [33].

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

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

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 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 one by one to the clients with the largest fractional remainders, so that $\sum_{k=1}^N c_k^{(i,j)} = n_{ij}$. Finally, for each class (i, j) we shuffle its n_{ij} sample indices and slice them into blocks of size $c_k^{(i,j)}$. Client k then collects all its four blocks across (i, j) , yielding a partition that in expectation realizes the desired within-client correlation γ_k between A and Y .

This approach can be regarded as a generalization of the synergy-level-based heterogeneous split in [33] to the multi-client setting, where the A - Y correlation for each client is governed by a parameter randomly drawn from $[a, b] \subseteq [0, 1]$, thereby yielding a more pronounced balance between global fairness and local fairness. Throughout the experimental evaluation, we set $\gamma_k \in [0.2, 0.8]$ to guarantee that every client has a sufficient number of sensitive group samples to assess local fairness.

D Detailed Experiments Results

D.1 Comparison Result and binary EOP criterion

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

The Pareto curve for the global DP criterion is shown in Figure 2 where the horizontal axis denotes accuracy and the vertical axis represents the fairness metric. Consequently, models located closer to the upper-right corner exhibit superior accuracy-fairness trade-offs. As illustrated in Figure 2, our method outperforms all existing state-of-the-art approaches when comparing accuracy against either global fairness in isolation.

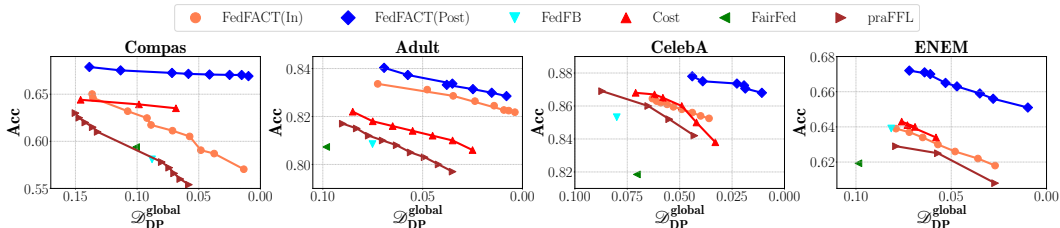


Figure 2: The Pareto frontier on Compas, Adult, CelebA and ENEM datasets. The curve closer to the upper right corner indicates a better trade-off between accuracy and fairness.

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

Parato Curves of EOP. In Figure 3, we illustrate the Pareto curve for the Equalized Odds (EO) criterion-accuracy. Because EOP enforces tighter constraints than DP, precise adherence in a federated context requires large per-group sample counts at each client. Hence, we also compare the global EOP here. Our framework still exceeds all state-of-the-art baselines in trading off accuracy against fairness.

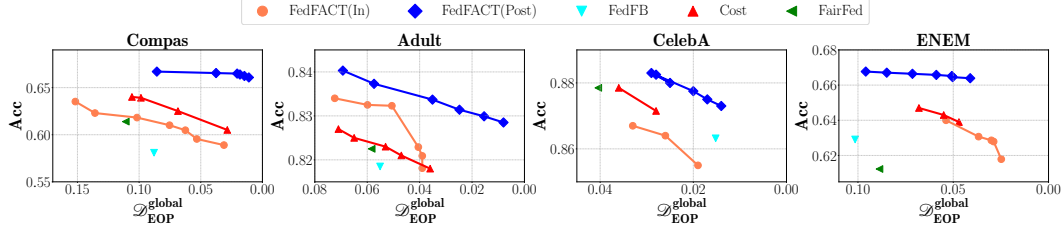


Figure 3: The Pareto frontier on Compas, Adult, CelebA and ENEM datasets. The curve closer to the upper right corner indicates a better trade-off between accuracy and fairness.

D.2 Details for Multi-Class Classification

Multi-Class fair datasets. We illustrate how FedFACT performs on multi-class prediction using CelebA and ENEM. For CelebA, with ‘Gender’ still serving as the sensitive attribute, We employ the binary attributes ‘Smile’ and ‘Big_Nose’ to construct a multiclass task by mapping their joint values $\{0, 1\} \times \{0, 1\}$ onto a four-class label set $\{0, 1, 2, 3\}$, thereby formulating a multiclass classification problem on the CelebA dataset. These attributes are commonly used in centralized machine learning literature [13, 97] to construct fairness-aware classification tasks. For ENEM, we follow [3] to quantize the Humanities exam score to 5 classes. In order to guarantee adequate per-group sample sizes at each client in heterogeneous settings for fairness evaluation (or some clients only hold less than 10 samples for specific group under heterogeneous partitioning), we adopt the four race labels ‘Branca,’ ‘Preta,’ ‘Parda,’ and ‘Amarela’ from the Race attribute as the sensitive groups. These datasets are partitioned into five clients under a heterogeneous split with $\gamma = 1$.

Evaluation. In terms of baselines, only the Cost [25] algorithm is theoretically applicable to fairness optimization in multiclass federated learning scenarios. However, their experiments and code are limited to binary classification, and have already been used as binary baselines for comparison with our method. Consequently, we focus exclusively on reporting FedFACT’s performance along with FedAvg in multiclass fairness, establishing it as a pioneering approach in this setting.

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	$\mathcal{D}^{global}$	$\mathcal{D}^{local}$	Acc	$\mathcal{D}^{global}$	$\mathcal{D}^{local}$	Acc	$\mathcal{D}^{global}$	$\mathcal{D}^{local}$	Acc	$\mathcal{D}^{global}$	$\mathcal{D}^{local}$
(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	$\mathcal{D}^{global}$	$\mathcal{D}^{local}$	Acc	$\mathcal{D}^{global}$	$\mathcal{D}^{local}$	Acc	$\mathcal{D}^{global}$	$\mathcal{D}^{local}$	Acc	$\mathcal{D}^{global}$	$\mathcal{D}^{local}$
1	61.17	0.0407	0.0732	82.04	0.0014	0.0401	86.15	0.0382	0.0473	65.33	0.0293	0.0387
2	61.29	0.0408	0.0731	81.24	0.0015	0.0416	85.54	0.0382	0.0482	65.54	0.0285	0.0392
5	61.18	0.0410	0.0723	81.63	0.0032	0.0397	85.91	0.0377	0.0472	65.41	0.0307	0.0390
10	61.14	0.0404	0.0736	81.91	0.0048	0.0399	86.59	0.0384	0.0471	65.11	0.0398	0.0383

D.5 Efficiency and Scalability Study

In this section, we conduct out 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.

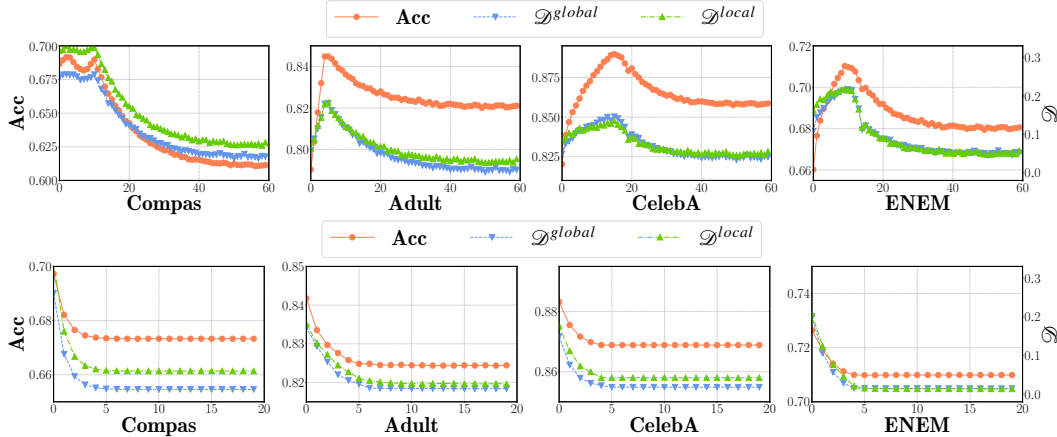


Figure 4: Communication Effectiveness Analysis. The convergence rates of both the in-processing (top row) and post-processing (bottom row) methods with respect to communication rounds on Compas, Adult, CelebA, and ENEM datasets.

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 commonly used DP, EOP and EOP criteria, but it excludes some nonlinear formulations of fairness, e.g. Predictive Parity [22] and individual fairness [29]. 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 characteristics to optimize fairness may reduce the sample complexity needed for local fairness optimization. Addressing these limitations remains an important avenue for future work.

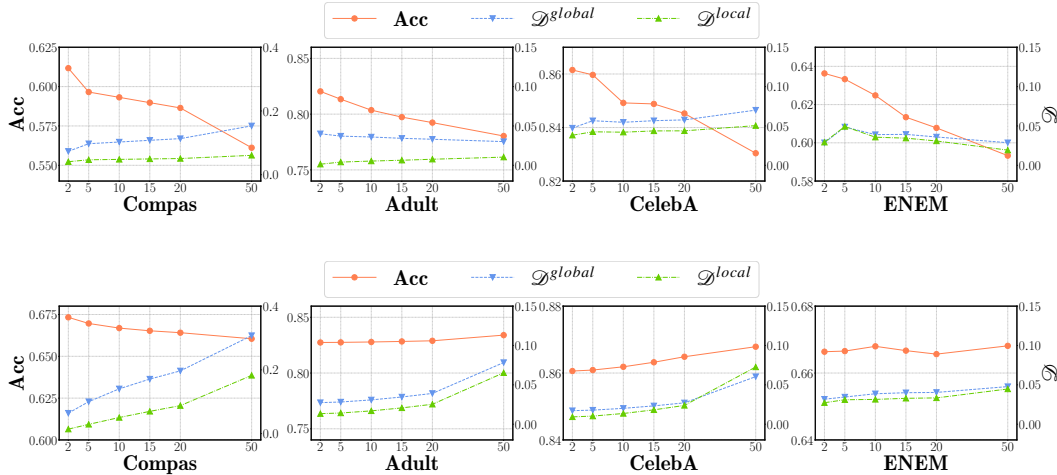


Figure 5: Scalability Analysis. The behavior of both the in-processing (top row) and post-processing (bottom row) methods as the number of clients increases from 2 to 50 across Compas, Adult, CelebA, and ENEM datasets.