# GLOBALLY COORDINATED SAMPLE REWEIGHTING FOR FAIR FEDERATED LEARNING UNDER HETEROGENEITY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Federated learning (FL) enables collaborative model training across decentralized institutions, making group fairness an important requirement when the resulting global model is deployed across demographically heterogeneous users. However, assessing global group fairness requires statistics aggregated across clients, whereas FL updates the model using private local data. This inherent mismatch makes it challenging to guarantee global group fairness, especially under client heterogeneity. To address this problem, we introduce *GloReFair*, a globally coordinated sample reweighting framework that bridges this gap by converting global fairness feedback into local training signals. The key idea is to separate where fairness correction is needed from how training emphasis should be allocated. We realize this separation through a server-side controller that adjusts channel prices using global fairness residuals. Given these prices, clients compute sample weights using a unique closed-form rule derived from the Lagrangian of an idealized round-wise allocation problem. We further establish theoretical guarantees on centralized equivalence and strict controller expressivity. Across three federated benchmarks, GloReFair achieves favorable fairness-utility tradeoffs and remains robust as client heterogeneity becomes severe, including settings in which competing fair-FL methods substantially degrade.

## 1 INTRODUCTION

Federated learning (FL) trains a shared model from decentralized data by alternating client-side optimization and server-side aggregation (McMahan et al., 2017). While this paradigm keeps raw training data on clients, it introduces challenges to group fairness protection. In detail, global group fairness is assessed using statistics pooled across clients, while model updates are computed on local data without direct access to the joint demographic distribution across clients. Thus, client-level fairness constraints cannot effectively guarantee the global group fairness (Ezzeldin et al., 2023; Wang et al., 2024). And the mismatch is exacerbated under client heterogeneity, where demographic compositions and scales vary substantially across each other. This motivates our core problem: how to align local model updates with global group fairness objective under client heterogeneity?

Some approaches have been proposed to address group fairness at different levels, such as local debiasing methods (Abay et al., 2020), FedFB (Zeng et al., 2022) and FairFed (Ezzeldin et al., 2023). Local debiasing can be distorted by skewed or incomplete demographic support at individual clients. FedFB coordinates local training through group-label coefficients updated using global feedback, but it applies a common reweighting coefficient to samples within each group–label subset. FairFed adjusts client contributions during model aggregation. These observations motivate a mechanism that converts global fairness feedback into local training signals that target specific group-label subsets while adapting to sample difficulties across heterogeneous clients. This issue is particularly relevant to label-conditional fairness criteria, such as equal opportunity (EO) and equalized odds (EOD), for which violations act through different group-label subsets.

To handle this challenge, we consider using global fairness feedback for coordinating local training: the server identifies which group-label channels require more attention, while each client determines the contribution of individual samples within those channels according to their difficulties. Here, a



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

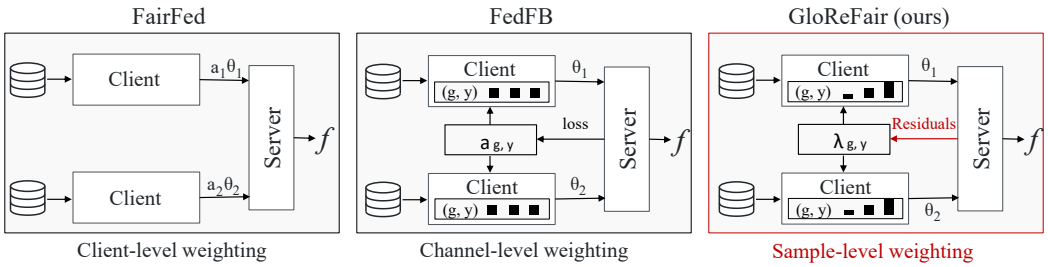


Figure 1: The architecture comparison between FairFed, FedFB and GloReFair.

group-label channel consists of samples sharing the same sensitive group and target label. Based on this principle, we propose *GloReFair*, a globally coordinated sample reweighting framework for fair FL. In the proposed framework, the server maintains one scalar control parameter for each channel, termed its channel price, and updates it using global fairness residuals. These prices provide a common coordination signal across clients and are combined with local sample losses to re-weight samples within each channel. As a result, the mechanism separates two complementary tasks: determining where fairness correction is needed and how to distribute training emphasis. Figure 1 shows the architecture difference between FairFed, FedFB and GloReFair.

Different from the existing fair FL, this design is motivated from an attention allocation problem and therefore admits a simple optimization interpretation. We introduce an idealized round-wise allocation problem that equalizes total training attention across active groups within each label class. For fixed centered channel prices, maximizing its Lagrangian over nonnegative sample weights yields a unique closed-form weighting rule. By leveraging global fairness feedback and fine-grained sample reweighting, GloReFair matches the state-of-the-art (SOTA) approaches on accuracy and fairness, and it offers a significant advantage in terms of resistance to client heterogeneity. To summarize, our main contributions are as follows:

- We propose a globally coordinated sample reweighting framework for fair FL. Different from existing approaches, the proposed framework employs global fairness residuals for sample-specific training correction, which helps coordinate local training across heterogeneous clients to reduce global disparities.
- We provide a theoretical characterization of the underlying mechanisms. Under full client participation, we establish partition invariance of global fairness feedback and synchronized sample reweighting, and show that one full-batch local step recovers the corresponding centralized weighted-gradient update. We further establish the expressivity advantage of label-group control, giving boundedness and regret guarantees for the controller.
- We evaluate the proposed method on three standard FL benchmarks under different client heterogeneity. Across these settings, it achieves favorable fairness-utility trade-off, particularly under strong client heterogeneity.

## 2 Related Work

**Federated optimization and heterogeneity.** Federated Averaging (FedAvg) builds the standard federated optimization paradigm of alternating client-side stochastic-gradient updates with server-side aggregation (McMahan et al., 2017). Although client-side training substantially reduces the server-client communication frequency, heterogeneous client data can cause successive drifts of local updates, thereby impairing global optimization. Therefore, substantial methodologies have been studied to address client heterogeneity. For example, FedProx regularizes each local objective toward the current global model to improve optimization stability (Li et al., 2020). SCAFFOLD introduces control variates to correct client drift induced by heterogeneous local data (Karimireddy et al., 2020). Other classical methods include FedNova (Wang et al., 2020) and FedDyn (Acar et al., 2021). Built on the standard FL paradigm, GloReFair also employs proximal regularization to stabilize local optimization, but targets on global fairness protection.

**Group fairness in federated learning.** Existing fair FL methods apply fairness corrections through different mechanisms. Aggregation-based approaches, including FairFed (Ezzeldin et al., 2023) and FODA (Liu et al., 2026), adjust client contributions during server-side model aggregation. Other approaches incorporate fairness directly into the learning objective, such as Bounded Group Loss (BGL) (Hu et al., 2024), FairTrade (Badar et al., 2024) and Function Tracking (Rychener et al., 2025). Other methods address fairness beyond the training process, as represented by LoGoFair (Zhang et al., 2025a) and FedFACT (Zhang et al., 2025b) and OursFed (Xin et al., 2026). A closely related approach is FedFB, which coordinates local training through group-label reweighting coefficients updated from globally aggregated statistics (Zeng et al., 2022). GloReFair also uses global feedback to guide local reweighting, but combines shared channel prices with individual sample losses to determine sample-specific weights within each group-label channel.

**Fair classification and adaptive reweighting.** EO and EOD measure disparities in label-conditional prediction rates across groups (Hardt et al., 2016). Centralized fair classification incorporates these constraints into model via multiple paradigms (Agarwal et al., 2018; Cotter et al., 2019). Adaptive reweighting methods modify training emphasis in response to fairness signals, so they are closely related to our approach. For example, Chai and Wang learn sample-specific weights that emphasize error-prone samples while maintaining balance across demographic groups (Chai & Wang, 2022). Intrinsic reweighting is another method developed for the worst-case fairness objective (Li et al., 2025). Following (Chai & Wang, 2022), GloReFair shares the threshold-based weighting form and addresses its coordination across heterogeneous clients. In our framework, the thresholds are shared group-label prices updated from globally pooled EO/EOD residuals, while sample losses remain locally evaluated. This design connects local training emphasis to disparities measured on the pooled population, including when individual clients lack some globally present channels. Relative to FedFB's global group-label coordination, GloReFair additionally differentiates sample weights within each channel according to individual losses. These distinctions motivate our study of residual-driven sample reweighting under heterogeneous federated partitions.

## 3 PRELIMINARIES

**Federated binary classification.** We consider $K$ clients and all of them participate the FL. Client $k$ holds a local dataset $\mathcal{D}_k = \{(x_{k,i}, y_{k,i}, s_{k,i})\}_{i=1}^{n_k}$, where $y_{k,i} \in \{0, 1\}$ is the target label, $s_{k,i} \in \{1, \ldots, G\}$ is the sensitive group membership, and $n_k$ is the number of local samples. Model $f_\theta(x)$ outputs the probability of $Y = 1$ and each sample incurs binary cross-entropy loss $\ell_{k,i}(\theta)$. With $N = \sum_k n_k, p_k = n_k/N$, the empirical risk minimization objective is:

$$F(\theta) = \sum_{k=1}^{K} p_k F_k(\theta), \qquad F_k(\theta) = \frac{1}{n_k} \sum_{i=1}^{n_k} \ell_{k,i}(\theta).$$

**EO and EOD metrics.** Following (Hardt et al., 2016; Weerts et al., 2023), we use empirical EO and EOD metrics. In the method and analysis, all fairness rates are computed on the pooled training data. Let $n_{g,y}$ be the global sample number in channel $(g, y)$, the active set for label $y$ is $A_y = \{g : n_{g,y} > 0\}$. For a classifier with prediction $\widehat{Y}$, group-specific True Positive Rate (TPR) and False Positive Rate (FPR) are:

$$\mathrm{TPR}_g(\theta) = \Pr(\widehat{Y} = 1 \mid S = g, Y = 1), \quad g \in A_1,$$

$$\mathrm{FPR}_g(\theta) = \Pr(\widehat{Y} = 1 \mid S = g, Y = 0), \quad g \in A_0.$$

The EO gap is used to measure positive-label disparity:

$$\Delta_{\mathrm{EO}}(\theta) = \max_{g \in A_1} \mathrm{TPR}_g(\theta) - \min_{g \in A_1} \mathrm{TPR}_g(\theta).$$

And the EOD gap additionally accounts for the disparity on the negative-label slice:

$$\Delta_{\mathrm{EOD}}(\theta) = \max\{\Delta_{\mathrm{EO}}(\theta), \max_{g \in A_0} \mathrm{FPR}_g(\theta) - \min_{g \in A_0} \mathrm{FPR}_g(\theta)\}.$$

**Residual signals.** Let $\overline{\mathrm{TPR}}_{A_1}$ and $\overline{\mathrm{FPR}}_{A_0}$ be the arithmetic means of the group-specific rates over their respective active sets. We define the hard fairness residuals as

$$c_g^+(\theta) = \overline{\mathrm{TPR}}_{A_1}(\theta) - \mathrm{TPR}_g(\theta), \qquad c_g^-(\theta) = \mathrm{FPR}_g(\theta) - \overline{\mathrm{FPR}}_{A_0}(\theta).$$

Thus, a positive $c_g^+$ indicates that group $g$ has a below-average TPR and a positive $c_g^-$ indicates an above-average FPR. We further introduce score-based counterparts for each group-label channel:

$$\tau_g(\theta) = \frac{1}{n_{g,1}} \sum_{k,i:s_{k,i}=g,y_{k,i}=1} f_\theta(x_{k,i}), \qquad \phi_g(\theta) = \frac{1}{n_{g,0}} \sum_{k,i:s_{k,i}=g,y_{k,i}=0} f_\theta(x_{k,i}),$$

Similarly, let $\bar{\tau}_{A_1}$ and $\bar{\phi}_{A_0}$ denote their arithmetic means over the corresponding active groups. Then, the smooth fairness residuals are:

$$\widetilde{c}_g^+(\theta) = \bar{\tau}_{A_1}(\theta) - \tau_g(\theta), \qquad \widetilde{c}_g^-(\theta) = \phi_g(\theta) - \bar{\phi}_{A_0}(\theta).$$

These residuals serve as surrogates for the hard because they vary continuously with model scores and provide differentiable feedback. By construction, the hard residuals are zero-sum within each label slice:

$$\sum_{g\in A_1} c_g^+(\theta) = 0, \qquad \sum_{g\in A_0} c_g^-(\theta) = 0.$$

and the same property holds for $\widetilde{c}_g^+$ and $\widetilde{c}_g^-$. For EOD, the fairness-aware learning objective can be conceptualized as

$$\begin{aligned} \min_{\theta} \quad & F(\theta) \\ \text{s.t.} \quad & c_g^+(\theta) \le 0, \quad g \in A_1, \\ & c_g^-(\theta) \le 0, \quad g \in A_0. \end{aligned} \tag{1}$$

Obviously, satisfying every inequality forces parity within the corresponding label slice because the residuals are zero-sum within each label slice. For EO criterion, the learning objective retains only the positive-label constraints.

Problem (1) describes the fairness objective on the pooled data. However, it is not directly amenable to FL because no individual client can evaluate global residuals from its local data alone. Moreover, the hard residuals are piecewise constants of sample scores for a fixed classification threshold. GloReFair therefore combines the hard residuals with their smooth counterparts for coordinating fairness-aware local optimization.

## 4 Methodology and Analysis

In this section, we introduce the core components of GloReFair with theoretical properties motivating its design.

### Allocation-Derived Local Training

At communication round $t$, let $\theta^t$ be the synchronized model and $\ell_i^t = \ell_i(\theta^t)$ be the corresponding loss of sample $i$. We consider an idealized attention allocation problem over the $N$ samples: each sample is associated with a nonnegative weight $w_i$, the total attention assigned to channel $(g, y)$ is $W_{g,y}(w) = \sum_{i:s_i=g,y_i=y} w_i$. With $m_y = |A_y|$ and $W_y = \sum_{g\in A_y} W_{g,y}$, we want:

$$\begin{aligned} \max_{w\ge 0} \quad & \sum_{i=1}^{N} \left(w_i \ell_i^t - \alpha w_i^2\right) \\ \text{s.t.} \quad & W_{g,y}(w) = \frac{W_y(w)}{m_y}, \quad g \in A_y,\ y \in \{0,1\}, \end{aligned} \tag{2}$$

where $\alpha > 0$ is the quadratic allocation-penalty coefficient which controls the overall attention scale. The constraints equalize attention assigned to the active channels within each label slice, while leaving the overall attention $W_y(w)$ endogenous. We denote the resulting centered price of active channel $(g, y)$ as $\lambda_{g,y}^t$, with $\sum_{g\in A_y} \lambda_{g,y}^t = 0$. For fixed prices, the Lagrangian is separable

across samples and strictly concave in each $w_i$. Maximizing each coordinate yields the unique sample-wise maximizer:

$$w_i^{t,\star} = \left[\frac{\ell_i^t - \lambda_{s_i,y_i}^t}{2\alpha}\right]_+. \tag{3}$$

Equation (3) reveals the pricing interpretation: the sample weight increases with its loss and decreases with the corresponding channel price when $\ell_i^t > \lambda_{s_i,y_i}^t$. GloReFair adopts this parameterization and updates the prices using global fairness residuals. These residual-driven updates are not dual optimization steps for Problem (2) and do not enforce its attention-balance constraints.

At round $t$, client $k$ receives $(\theta^t, \Lambda^t)$ from the server, evaluates sample losses under $\theta^t$, computes the corresponding weights via (3), and freezes the weights while approximately minimizing

$$J_{k,t}^{\mathrm{fr}}(\theta) = \frac{1}{n_k}\sum_{i=1}^{n_k} w_{k,i}^{t,\star}\ell_{k,i}(\theta) + \frac{\mu}{2}\|\theta - \theta^t\|_2^2. \tag{4}$$

If each sample loss $\ell_{k,i}$ is finite and convex on $\mathbb{R}^d$ and $\mu > 0$, then (4) is $\mu$-strongly convex and has a unique minimizer. The same procedures are also applicable to neural networks (NN), but the proximal term cannot guarantee the uniqueness of the local minimizer. Appendix A provides the allocation derivation and local-target analysis. With the controller fixed and bounded Lipschitz sample losses along the local trajectory, Appendix C bounds the frozen-weight objective mismatch at the same model parameters in terms of local model drift.

### Residual-Driven Price Control

The server computes the global empirical hard and smooth residuals by aggregating channel counts, positive-prediction counts, and score sums across all clients, and forms the feedback signals as:

$$r_g^+ = \beta_t c_g^+ + (1-\beta_t)\widetilde{c}_g^+, \qquad r_g^- = \beta_t c_g^- + (1-\beta_t)\widetilde{c}_g^-, \tag{5}$$

where all residuals are evaluated at $\theta^{t+1}$ and $\beta_t \in [0,1]$. The server computes channel rates from the aggregated counts and sums; a locally absent channel contributes zero statistics. Under EOD, the controller updates both label slices using $r_g^+$ and $r_g^-$, respectively. Under EO, only the positive-label prices are updated. The negative-label prices are initialized at zero and remain fixed at $\lambda_{g,0}^t = 0$ for every group and round.

The controller is constrained to the feasible set

$$\mathcal{C} = \left\{ \Lambda : \begin{array}{ll} |\lambda_{g,y}| \le \lambda_{\max}, & \\ \lambda_{g,y} = 0 & (g \notin A_y), \\ \sum_{g\in A_y} \lambda_{g,y} = 0 & (y \in \{0,1\}) \end{array} \right\}. \tag{6}$$

Let $\gamma_t = (1+\rho u_t)^{-1/2}$ be the attenuation coefficient, where $\rho > 0$ and $u_t$ denotes the number of controller updates completed before the controller update in round $t$, with $u_t = 0$ at the first controller update. $\eta_{\mathrm{tpr}} > 0$ and $\eta_{\mathrm{fpr}} > 0$ denote step sizes of price adjustment under EOD, while $\eta_{\mathrm{tpr}} > 0$ and $\eta_{\mathrm{fpr}} = 0$ under EO. Then, the update is

$$\begin{aligned} v_{g,1}^{t+1} &= \lambda_{g,1}^t - \gamma_t\eta_{\mathrm{tpr}} r_g^+, \qquad v_{g,0}^{t+1} = \lambda_{g,0}^t - \gamma_t\eta_{\mathrm{fpr}} r_g^-, \\ \lambda_{g,y}^{t+1} &= \mathrm{clip}(v_{g,y}^{t+1} - \zeta_y, -\lambda_{\max}, \lambda_{\max}), \end{aligned} \tag{7}$$

where $\zeta_y$ is chosen to keep the active coordinates zero-sum. Inactive coordinates are also fixed at zero. This form gives the exact Euclidean projection onto the centered box-constrained slice; its computation is provided in Appendix F.

At the fixed model $\theta^{t+1}$, whenever the raw price update is feasible, the projection leaves it unchanged. A positive TPR deficit ($r_g^+ > 0$) lowers $\lambda_{g,1}$ and increases the weights of positive-label samples in group $g$ when their losses exceed the updated channel price; under EOD, a positive FPR excess ($r_g^- > 0$) has the same effect on negative-label samples. Under local logit concentration and non-saturation at the label-specific reference logits, Appendix D relates smooth fairness residuals to channel loss deviations: their signs agree when the first-order terms dominate the remainders. For

hybrid signals ($0 < \beta_t < 1$), agreement between the hard and smooth residual signs is a sufficient condition for the mixture to retain the sign of its smooth counterpart. Algorithm 1 presents one communication round of GloReFair.

**Algorithm 1** Procedures of GloReFair in one communication round

**Require:** synchronized model $\theta^t$, controller $\Lambda^t$, target criterion (EO or EOD)

1: Server broadcasts $(\theta^t, \Lambda^t)$ to all the clients.
2: Each client $k$ computes frozen weights using (3) and performs local SGD steps on (4) in parallel.
3: Aggregate $\theta^{t+1} = \sum_{k=1}^{K} p_k \theta_k^{t+1}$ and broadcast $\theta^{t+1}$ to all the clients.
4: Each client evaluates predictions under $\theta^{t+1}$ and returns group-label counts, positive-prediction counts, and score sums.
5: Aggregate the returned statistics, obtain feedback signals through (5).
6: Update and project $\Lambda^{t+1}$ using (7).
7: **return** $(\theta^{t+1}, \Lambda^{t+1})$

### THEORETICAL PROPERTIES

We analyze the theoretical properties of GloReFair to explain the roles of shared prices, label-specific control, and projected updates. These properties include single-step centralized equivalence, strict controller expressivity, and controller boundedness and online regret. We first show that the shared prices yield an exact aggregation identity at the synchronized starting point.

**Proposition 1** (Single-step centralized equivalence)**.** *Fix $(\theta^t, \Lambda^t)$, and suppose all sample losses are differentiable at $\theta^t$. Suppose every client starts from $\theta^t$ and takes exactly one full-batch gradient-descent step on* (4)*, using the same step size $\eta > 0$. With full participation and aggregation weights $p_k = n_k/N$, the resulting model satisfies*

$$\theta^{t+1} = \theta^t - \frac{\eta}{N} \sum_{k=1}^{K} \sum_{i=1}^{n_k} w_{k,i}^{t,\star} \nabla \ell_{k,i}(\theta^t).$$

*For fixed global data and $(\theta^t, \Lambda^t)$, this update is invariant to repartitioning the samples among nonempty clients.*

Appendix B provides the proof and establishes partition invariance of global fairness feedback at a fixed common model, classification rule, and mixture coefficient. This equivalence concerns a single full-batch step from a common model state.

Proposition 2 shows the group-label channel pricing is more expressive than the group-specific.

**Proposition 2** (Strict controller expressivity)**.** *Under EOD criterion, suppose $A_0 = A_1 = A$ and $\lambda_{\max} \geq 0$, and define $\mathcal{C}_{\text{coarse}} = \{\Lambda \in \mathcal{C} : \lambda_{g,0} = \lambda_{g,1}$ for every $g \in A\}$. For any direction $D$ supported on $A \times \{0, 1\}$ and centered within each label slice,*

$$\max_{\Lambda \in \mathcal{C}} \langle \Lambda, D \rangle \geq \max_{\Lambda \in \mathcal{C}_{\text{coarse}}} \langle \Lambda, D \rangle.$$

*If $|A| \geq 2$ and $\lambda_{\max} > 0$, there exists such a direction $D$ for which the inequality is strict.*

Appendix E provides the proof and a construction with opposing label-wise directions. The comparison concerns the attainable linear controller reward at a fixed direction.

Finally, we analyze the boundedness and online regret of the projected controller. Let $q$ index controller updates and define the scaled residual vector $\widetilde{d}^q \in \mathbb{R}^{2G}$, with active-channel components $\widetilde{d}_{g,1}^q = -\eta_{\text{tpr}} r_g^{+,q}$ and $\widetilde{d}_{g,0}^q = -\eta_{\text{fpr}} r_g^{-,q}$, where $r_g^{\pm,q}$ denotes the feedback at update $q$ and inactive components are zero.

**Proposition 3** (Controller boundedness and online regret)**.** *Assume that $\Lambda^0 \in \mathcal{C}$, the active sets remain fixed, no post-projection normalization is applied, and the label step sizes are fixed and finite. With $\gamma_q = (1 + \rho q)^{-1/2}$ and $\rho > 0$, the projected update* (7) *satisfies $\|\Lambda^q\|_2 \leq \lambda_{\max}\sqrt{2G}$ and, for any fixed $\Lambda \in \mathcal{C}$,*

$$\sum_{q=0}^{Q-1} \langle \Lambda - \Lambda^q, \widetilde{d}^q \rangle \leq O(\sqrt{Q}).$$

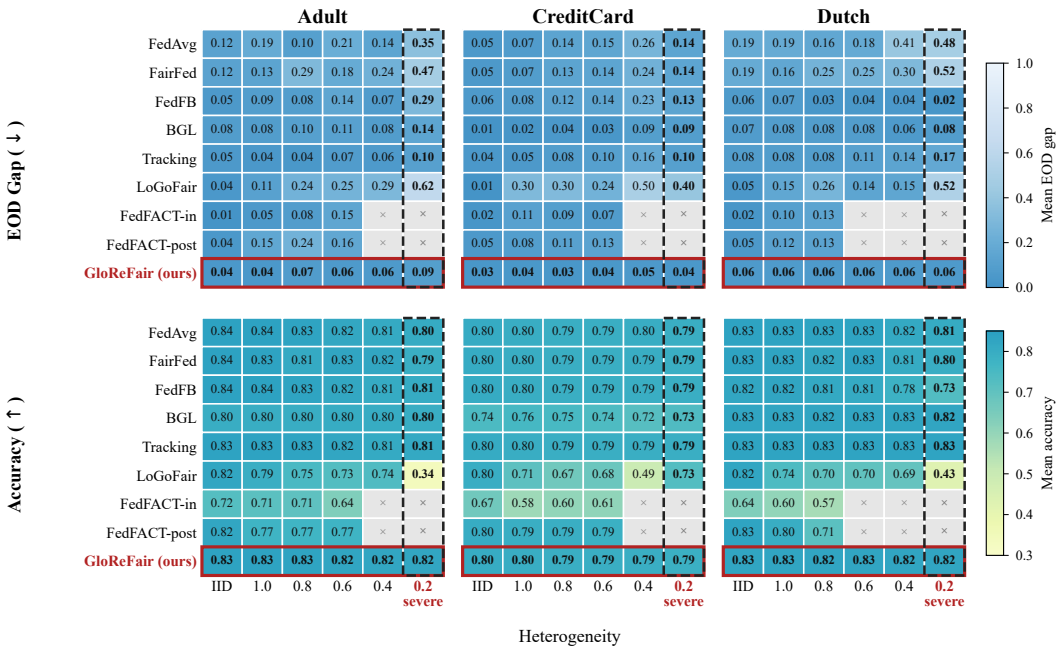


Figure 2: The comparison results under EOD and joint Dirichlet shifts. Darker cells indicate better performance. Grey crossed cells denote that the method does not work in the corresponding settings.

The residual construction gives bounded update directions; the explicit bound and proof are provided in Appendix F. The regret compares fixed feasible controllers on the same feedback sequence. Under the pure smooth feedback and additional model-comparison conditions, Appendix G translates this regret bound into a bound on the empirical smooth residuals of a score ensemble.

## 5 Experiments

This section demonstrates experiment results for method verification. Complete results are reported in the supplement.

### Experiments under EOD criterion

**Setup** We use three public benchmarks, including Adult (Kohavi, 1996), Default Credit Card (Yeh & Lien, 2009) and Dutch (Kang et al., 2022), with sex as the sensitive attribute. The client heterogeneity contains label, group and joint Dirichlet with 5 levels ($H \in \{1.0, 0.8, 0.6, 0.4, 0.2\}$). Lower $H$ indicates stronger heterogeneity. We use LR as the classifier embedded in each FL method and conduct hyper-parameter selection under independent and identically distributed (IID) assumption. All the reported values are means over 3 seeds. Optimization settings, prediction rule, and hyperparameter selection protocol are provided in Appendix H.

**Baselines** The comparison baselines include 8 state-of-the-art (SOTA) methods: FedAvg (McMahan et al., 2017), FedFB (Zeng et al., 2022), FairFed (Ezzeldin et al., 2023), BGL (Hu et al., 2024), Function Tracking (Rychener et al., 2025), the in-processing and post-processing variants of FedFACT (Zhang et al., 2025b), and LoGoFair (Zhang et al., 2025a).

**Results** As shown in Figure 2, we see that GloReFair achieves competitive performances in the IID and scenarios with mild heterogeneity. Figure 3 demonstrates the methods' performance under the severest joint heterogeneity. Note that in some scenarios with strong heterogeneity, FedFACT-in and FedFACT fail to work for violating their positive-cell-mass requirements. GloReFair is significantly

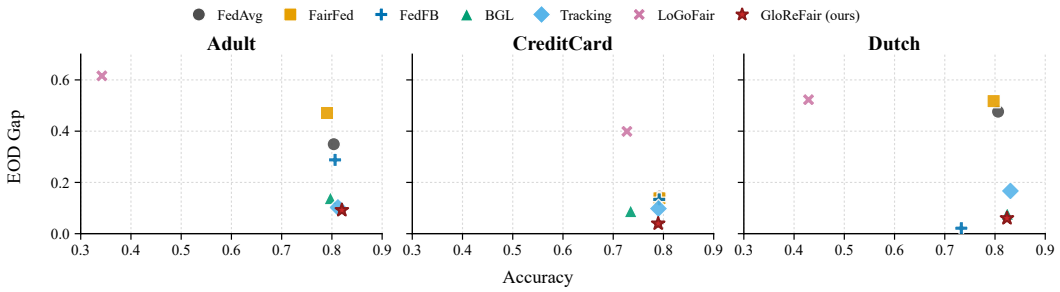


Figure 3: The performance under the severest joint Dirichlet shifts. Note: FedFACT-in and FedFACT-post are excluded because they can not work in the setting.

Table 1: Ablation under EOD criterion and joint Dirichlet shift ($H = 0.2$). Entries are means over three seeds, with standard deviations in parentheses.

| Variant | Adult | | CreditCard | | Dutch | |
|---|---|---|---|---|---|---|
| | Accuracy | EOD Gap | Accuracy | EOD Gap | Accuracy | EOD Gap |
| GloReFair | 0.820 (0.008) | **0.092 (0.050)** | **0.789 (0.005)** | **0.039 (0.009)** | 0.824 (0.002) | **0.060 (0.005)** |
| Local-ratio | **0.821 (0.011)** | 0.096 (0.074) | 0.788 (0.005) | 0.042 (0.009) | 0.822 (0.002) | 0.094 (0.033) |
| Coarse | 0.819 (0.016) | 0.172 (0.108) | 0.788 (0.006) | 0.115 (0.081) | 0.831 (0.001) | 0.342 (0.007) |
| Frozen-$\Lambda$ | 0.821 (0.014) | 0.158 (0.092) | 0.785 (0.007) | 0.098 (0.065) | **0.831 (0.002)** | 0.284 (0.019) |

more stable compared with the SOTA baselines. Full numerical results in terms of all the Dirichlet shifts are reported in Tables 3–5 in the appendix.

**Ablation Experiments** We set three variants related to GloReFair: Local-ratio replaces pooled group-label rates with equal-weight averages of client-local rates. Coarse ties the channel control within each group, and Frozen-$\Lambda$ isolate the contribution of adaptive price update. All the variants use the same data splits, model, hyper-parameters and random seeds. The results list in Table 1. Detailed explanation and result analysis are provided in Appendix H.

### Experiments under EO criterion

**Setup and baselines** GloReFair further uses the EO configuration described in Section 4, with zero negative-label prices and controller updates restricted to the positive-label slice. The used benchmarks, experimental settings and baselines keep the same as experiments under EOD criterion.

**Results** As shown in Figure 4, GloReFair also keep competitive performances in the scenarios with mild heterogeneity, and its superiority becomes obvious as the heterogeneity increases. Full numerical results in terms of all the Dirichlet shifts are reported in Tables 6–8 in the appendix.

**Ablation Experiments** Table 2 reports EO ablations under joint heterogeneity ($H = 0.2$), run separately from the comparison experiments. We omit Coarse because EO updates only the positive-label channel. See the result analysis in Appendix H.

## 6 Conclusion

In this paper, we introduce GloReFair for fair FL. This framework translates global fairness feedback into local training emphasis through a group-label channel controller. Across multiple benchmarks and heterogeneous settings, GloReFair reduces EOD or EO gaps while maintaining competitive

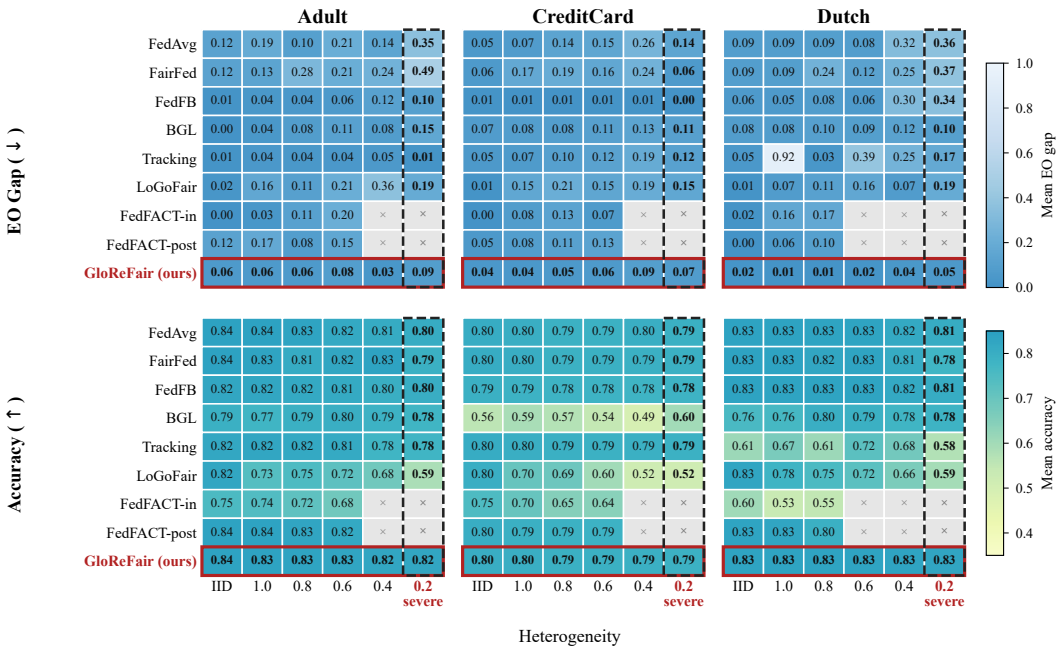


Figure 4: The comparison results under EO and joint Dirichlet shifts. Darker cells indicate better performance. Grey crossed cells denote that the method does not work in the corresponding settings.

Table 2: EO ablation under joint Dirichlet heterogeneity ($H = 0.2$). Entries are means over three seeds, with standard deviations in parentheses.

| Variant | Adult | | CreditCard | | Dutch | |
|---|---|---|---|---|---|---|
| | Accuracy | EO Gap | Accuracy | EO Gap | Accuracy | EO Gap |
| GloReFair | **0.821 (0.009)** | **0.088 (0.051)** | 0.791 (0.005) | **0.067 (0.029)** | 0.827 (0.002) | **0.048 (0.017)** |
| Local-ratio | 0.816 (0.017) | 0.187 (0.121) | 0.790 (0.004) | 0.069 (0.027) | 0.828 (0.004) | 0.059 (0.067) |
| Frozen-$\Lambda$ | 0.815 (0.018) | 0.213 (0.134) | **0.792 (0.007)** | 0.098 (0.065) | **0.831 (0.002)** | 0.176 (0.024) |

accuracy as client heterogeneity increases. These results suggest that GloReFair is a promising approach for fairness protection.

However, our work remains several limitations. First, we do not study intersectional fairness across multiple sensitive attributes or other fairness criteria, such as sufficiency. Second, although raw data remain at local clients, we establish no formal privacy or security guarantees for exchanged model updates and group-label statistics. Protection mechanisms such as secure aggregation and differential privacy are not evaluated. Third, we only study statistical heterogeneity in client data while assuming a shared model architecture and parameter space. Extending GloReFair to FL with heterogeneous client model architectures will make the method more applicable. We expect to discuss these problems in future work.

## AI Use Statement

Generative AI tools were used to assist in manuscript editing, formatting, and vibe coding. The authors assume responsibility for the final content of this work, including all text, claims and code.

### Ethics Statement

This work uses established research benchmarks for experiments and does not involve human subjects. So, no ethical approval is required.

### Reproducibility Statement

The appendix of this work includes the complete proofs of claims, implementation details, and experiment results. Access our code through `https://anonymous.4open.science/r/federated-fairness-2f4d9e7d/`.

## References


Annie Abay, Yi Zhou, Nathalie Baracaldo, Shashank Rajamoni, Ebube Chuba, and Heiko Ludwig. Mitigating bias in federated learning. *arXiv preprint arXiv:2012.02447*, 2020.

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul N. Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2111.04263.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 60–69. PMLR, 2018. URL https://proceedings.mlr.press/v80/agarwal18a.html.

Maryam Badar, Sandipan Sikdar, Wolfgang Nejdl, and Marco Fisichella. FairTrade: Achieving pareto-optimal trade-offs between balanced accuracy and fairness in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 10962–10970, 2024. doi: 10.1609/aaai.v38i10.28971. URL https://ojs.aaai.org/index.php/AAAI/article/view/28971.

Junyi Chai and Xiaoqian Wang. Fairness with adaptive weights. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2853–2866. PMLR, 2022. URL https://proceedings.mlr.press/v162/chai22a.html.

Andrew Cotter, Heinrich Jiang, Serena Wang, Taman Narayan, Maya Gupta, 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. URL http://jmlr.org/papers/v20/18-616.html.

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, pp. 7494–7502, 2023. doi: 10.1609/aaai.v37i6.25911. URL https://ojs.aaai.org/index.php/AAAI/article/view/25911.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, pp. 3315–3323, 2016. URL https://papers.nips.cc/paper_files/paper/2016/hash/6a9659feb1216f14f7384ba499518b38-Abstract.html.

Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. Fair federated learning via bounded group loss. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning*, pp. 140–160. IEEE, 2024. doi: 10.1109/SaTML59370.2024.00015. URL https://openreview.net/forum?id=YC824FKZex.

Jian Kang, Tiankai Xie, Xintao Wu, Ross Maciejewski, and Hanghang Tong. Infofair: Information-theoretic intersectional fairness. In *2022 IEEE International Conference on Big Data (Big Data)*, pp. 1455–1464. IEEE, 2022.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143. PMLR, 2020. URL https://proceedings.mlr.press/v119/karimireddy20a.html.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 202–207. AAAI Press, 1996.

Jing Li, Yinghua Yao, Yuangang Pan, Xuanqian Wang, Ivor W. Tsang, and Xiuju Fu. Alpha and prejudice: Improving $\alpha$-sized worst case fairness via intrinsic reweighting. *IEEE Transactions on Neural Networks and Learning Systems*, 36(10):18005–18019, 2025. doi: 10.1109/TNNLS.2025.3586609. URL https://doi.org/10.1109/TNNLS.2025.3586609.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, 2020. URL https://proceedings.mlsys.org/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html.

Ziyan Liu, Thomas A. Mazzuchi, and Shahram Sarkani. Fairness-optimized dynamic aggregation (FODA): A novel approach to equitable federated learning in heterogeneous environments. *Expert Systems with Applications*, 317:131807, 2026. doi: 10.1016/j.eswa.2026.131807. URL https://doi.org/10.1016/j.eswa.2026.131807.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Yves Rychener, Daniel Kuhn, and Yifan Hu. Global group fairness in federated learning via function tracking. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 865–873. PMLR, 2025. URL https://proceedings.mlr.press/v258/rychener25a.html.

Ganghua Wang, Ali Payani, Myungjin Lee, and Ramana Rao Kompella. Mitigating group bias in federated learning: Beyond local fairness. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=ANXoddnzct.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7611–7623, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/564127c03caab942e503ee6f810f54fd-Paper.pdf.

Hilde Weerts, Miroslav Dudik, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and improving fairness of ai systems. *Journal of Machine Learning Research*, 24(257):1–8, 2023. URL http://jmlr.org/papers/v24/23-0389.html.

Yun Xin, Jianfeng Lu, Gang Li, Shuqin Cao, Guanghui Wen, and Kehao Wang. OursFed: Provable group fairness-aware federated learning against distrust and fragility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27117–27125, 2026. doi: 10.1609/aaai.v40i32.39926. URL https://ojs.aaai.org/index.php/AAAI/article/view/39926.

I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009. doi: 10.1016/j.eswa.2007.12.020.

Yuchen Zeng, Hongxu Chen, and Kangwook Lee. Improving fairness via federated learning, 2022. URL https://arxiv.org/abs/2110.15545.

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, pp. 22470–22478, 2025a. doi: 10.1609/aaai.v39i21.34404. URL https://ojs.aaai.org/index.php/AAAI/article/view/34404.

Li Zhang, Zhongxuan Han, Xiaohua Feng, Jiaming Zhang, Yuyuan Li, and Chaochao Chen. FedFACT: A provable framework for controllable group-fairness calibration in federated learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/17ca2b635fdb94dbd3292da790e3fd46-Abstract-Conference.html.

# APPENDIX

## A ALLOCATION AND THE LOCAL TRAINING OBJECTIVE

We derive the centered channel prices and prove the local allocation and model-target claims in the main text. The allocation is computed at the synchronized model; the resulting weights are then held fixed during local training.

### A.1 CENTERED CHANNEL PRICES

At communication round $t$, the loss of sample $i$ is $\ell_i = \ell_i(\theta^t)$, which is a finite value. For non-empty label slices, let $A_y$ be the active groups and $m_y = |A_y|$. With $W_{g,y}(w) = \sum_{i:s_i=g,y_i=y} w_i$ and $W_y(w) = \sum_{g\in A_y} W_{g,y}(w)$, the idealized allocation problem is

$$\begin{aligned} \max_{w\geq 0} \quad & \sum_{i=1}^{N}(w_i\ell_i - \alpha w_i^2) \\ \text{s.t.} \quad & W_{g,y}(w) - \frac{W_y(w)}{m_y} = 0, \qquad g \in A_y. \end{aligned}$$

Here and below, sums over $y \in \{0,1\}$ omit empty label slices. Introducing unrestricted equality multipliers $\nu_{g,y}$ gives

$$\begin{aligned} \mathcal{L}(w,\nu) &= \sum_i (w_i\ell_i - \alpha w_i^2) - \sum_y \sum_{g\in A_y} \nu_{g,y}\left(W_{g,y} - \frac{W_y}{m_y}\right) \\ &= \sum_i (w_i\ell_i - \alpha w_i^2) - \sum_y \sum_{g\in A_y} (\nu_{g,y} - \bar{\nu}_y) W_{g,y}, \end{aligned}$$

where $\bar{\nu}_y = m_y^{-1}\sum_{h\in A_y}\nu_{h,y}$. Thus, the coefficient acting on each channel is its centered price

$$\lambda_{g,y} = \nu_{g,y} - \bar{\nu}_y, \qquad \sum_{g\in A_y}\lambda_{g,y} = 0.$$

Adding the same constant to all multipliers in a label slice leaves these prices unchanged. Conversely, any centered price vector is represented by taking $\nu_{g,y} = \lambda_{g,y}$. When $m_y = 1$, the balance constraint imposes no restriction and the corresponding price is zero.

GloReFair uses these centered prices as controller variables. Its updates rely on global fairness residuals rather than attention imbalance $W_{g,y} - W_y/m_y$, so it does not enforce the balance constraints above.

### A.2 OPTIMAL WEIGHTS FOR FIXED PRICES

For fixed prices, the Lagrangian separates across samples. On client $k$, the corresponding allocation objective is

$$\max_{w_k\geq 0} \frac{1}{n_k}\sum_{i=1}^{n_k}\left[w_{k,i}\big(\ell_{k,i}(\theta^t) - \lambda^t_{s_{k,i},y_{k,i}}\big) - \alpha w_{k,i}^2\right] \qquad (\alpha > 0).$$

Writing $a = \ell_{k,i}(\theta^t) - \lambda^t_{s_{k,i},y_{k,i}}$, each coordinate maximizes

$$aw - \alpha w^2 = -\alpha\left(w - \frac{a}{2\alpha}\right)^2 + \frac{a^2}{4\alpha}, \qquad w \geq 0.$$

The unique maximizer is $a/(2\alpha)$ when $a > 0$, and zero when $a \leq 0$. Hence the unique solution of the fixed-price allocation is

$$w^{t,\star}_{k,i} = \left[\frac{\ell_{k,i}(\theta^t) - \lambda^t_{s_{k,i},y_{k,i}}}{2\alpha}\right]_+. \qquad (8)$$

Strict concavity follows from the negative quadratic coefficient in every coordinate. A sample receives positive weight exactly when its loss exceeds its channel price. For fixed losses, increasing that price decreases the weight until it reaches zero.

### A.3 THE FROZEN-WEIGHT LOCAL OBJECTIVE

The weights in (8) are evaluated once at $\theta^t$. Holding them fixed gives the local training objective

$$J_{k,t}^{\mathrm{fr}}(\theta) = \frac{1}{n_k}\sum_i w_{k,i}^{t,\star}\ell_{k,i}(\theta) + \frac{\mu}{2}\|\theta-\theta^t\|_2^2, \qquad \mu > 0.$$

Suppose each loss is finite and convex on $\mathbb{R}^d$. Since the frozen weights are nonnegative, the weighted loss $H_{k,t}(\theta) = n_k^{-1}\sum_i w_{k,i}^{t,\star}\ell_{k,i}(\theta)$ is convex. Adding the proximal quadratic makes $J_{k,t}^{\mathrm{fr}}$ $\mu$-strongly convex, so it has at most one minimizer.

To establish existence, take any sub-gradient $v \in \partial H_{k,t}(\theta^t)$, which exists because $H_{k,t}$ is finite and convex on $\mathbb{R}^d$. The supporting affine bound yields

$$J_{k,t}^{\mathrm{fr}}(\theta) \geq H_{k,t}(\theta^t) + \langle v, \theta-\theta^t\rangle + \frac{\mu}{2}\|\theta-\theta^t\|_2^2.$$

The right-hand side tends to infinity as $\|\theta-\theta^t\|_2 \to \infty$. Thus the objective is coercive; it is also continuous and therefore attains its minimum on $\mathbb{R}^d$. Combined with strong convexity, it proves uniqueness.

The allocation formula requires only finite synchronized losses and $\alpha > 0$. The local-target uniqueness guarantee applies to convex parameterizations such as logistic regression when $\mu > 0$. For a general neural network, the proximal term alone cannot guarantee uniqueness of the local minimizer.

## B SINGLE-STEP CENTRALIZED EQUIVALENCE AND FEEDBACK INVARIANCE

The shared prices determine each sample's weight independently of client ownership. This gives the single-step centralized equivalence in the main text. Fix $(\theta^t, \Lambda^t)$, and define the pooled frozen-weight loss by

$$F_t^{\mathrm{fr}}(\theta) = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k} w_{k,i}^{t,\star}\ell_{k,i}(\theta),$$

where the weights are given by (8) and held fixed. Suppose all sample losses are differentiable at $\theta^t$. Every client starts from $\theta^t$ and takes a full-batch gradient descent step on $J_{k,t}^{\mathrm{fr}}$, with the same step size $\eta > 0$.

At the common starting point, the proximal term has gradient $\mu(\theta^t-\theta^t) = 0$. Therefore, the model returned by client $k$ is

$$\theta_k^{t+1} = \theta^t - \frac{\eta}{n_k}\sum_{i=1}^{n_k} w_{k,i}^{t,\star}\nabla\ell_{k,i}(\theta^t).$$

With full participation and aggregation weights $p_k = n_k/N$,

$$\theta^{t+1} = \sum_{k=1}^{K}\frac{n_k}{N}\theta_k^{t+1} = \theta^t - \frac{\eta}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k} w_{k,i}^{t,\star}\nabla\ell_{k,i}(\theta^t) = \theta^t - \eta\nabla F_t^{\mathrm{fr}}(\theta^t).$$

Thus, the aggregated model equals one gradient-descent step on the pooled frozen-weight loss. For fixed global data and $(\theta^t, \Lambda^t)$, repartitioning samples among nonempty clients does not change the sample weights and gradients.

**Lemma 1** (Partition invariance of global fairness feedback)**.** *Fix the pooled dataset, a common model $f_\theta$, a shared classification rule, and a mixture coefficient $\beta \in [0,1]$. Under full client participation, the global hard, smooth, and hybrid fairness residuals on each nonempty label slice are invariant to repartitioning the samples among clients.*

*Proof.* For client $k$, let $\mathcal{I}_{k,g,y} = \{i : s_{k,i} = g,\ y_{k,i} = y\}$. $\widehat{y}_{k,i}$ denotes the binary prediction obtained from the common model. Define the local channel statistics

$$n_{k,g,y} = |\mathcal{I}_{k,g,y}|, \qquad h_{k,g,y} = \sum_{i\in\mathcal{I}_{k,g,y}}\widehat{y}_{k,i}, \qquad a_{k,g,y} = \sum_{i\in\mathcal{I}_{k,g,y}} f_\theta(x_{k,i}).$$

When a channel is absent locally, all three statistics are zero, so no local channel rate needs to be defined. For every globally active channel, with $n_{g,y} = \sum_k n_{k,g,y} > 0$, the pooled rates are

$$R^{\mathrm{hard}}_{g,y} = \frac{\sum_k h_{k,g,y}}{\sum_k n_{k,g,y}}, \qquad R^{\mathrm{smooth}}_{g,y} = \frac{\sum_k a_{k,g,y}}{\sum_k n_{k,g,y}}.$$

For $y = 1$, these are $\mathrm{TPR}_g$ and $\tau_g$; for $y = 0$, they are $\mathrm{FPR}_g$ and $\phi_g$. Each numerator and denominator is a sum over the same pooled samples evaluated at the same model. Repartitioning therefore leaves the channel counts, active sets $A_y$, and pooled rates unchanged. Their arithmetic means over active groups are also unchanged. Consequently, the hard residuals $c^{\pm}_g$, smooth residuals $\widetilde{c}^{\pm}_g$, and mixtures $r^{\pm}_g = \beta c^{\pm}_g + (1-\beta)\widetilde{c}^{\pm}_g$ are invariant to the partition. □

## C FROZEN-WEIGHT STALENESS

We derive the frozen-weight staleness bound. Fix controller $\Lambda^t$ and recall the allocation rule

$$w_{k,i}(\theta; \Lambda^t) = \left[\frac{\ell_{k,i}(\theta) - \lambda^t_{s_{k,i},y_{k,i}}}{2\alpha}\right]_+, \qquad w^{t,\star}_{k,i} = w_{k,i}(\theta^t; \Lambda^t).$$

The client trains with the frozen weights $w^{t,\star}_{k,i}$. To measure the effect of freezing, define a comparison function that recomputes the weights at the parameter $\theta$ being evaluated:

$$J^{\mathrm{dyn}}_{k,t}(\theta) = \frac{1}{n_k}\sum_i w_{k,i}(\theta; \Lambda^t)\ell_{k,i}(\theta) + \frac{\mu}{2}\|\theta - \theta^t\|_2^2,$$
$$J^{\mathrm{fr}}_{k,t}(\theta) = \frac{1}{n_k}\sum_i w^{t,\star}_{k,i}\ell_{k,i}(\theta) + \frac{\mu}{2}\|\theta - \theta^t\|_2^2.$$

Both functions evaluate the losses at the same $\theta$; they differ only in whether the weights are computed at $\theta$ or $\theta^t$. Let $\mathcal{T}_{k,t}$ be the local training trajectory, including its starting point $\theta^t$. Assume that for every sample $i$ and all $\theta, \theta' \in \mathcal{T}_{k,t}$,

$$0 \le \ell_{k,i}(\theta) \le B, \qquad |\ell_{k,i}(\theta) - \ell_{k,i}(\theta')| \le L\|\theta - \theta'\|_2.$$

Since the positive-part map is 1-Lipschitz and the channel price is fixed, for any $\theta \in \mathcal{T}_{k,t}$,

$$|w_{k,i}(\theta; \Lambda^t) - w^{t,\star}_{k,i}| \le \frac{|\ell_{k,i}(\theta) - \ell_{k,i}(\theta^t)|}{2\alpha} \le \frac{L}{2\alpha}\|\theta - \theta^t\|_2.$$

This inequality also holds when a weight changes between zero and a positive value. The identical proximal terms cancel in the objective difference, giving

$$\begin{aligned} |J^{\mathrm{dyn}}_{k,t}(\theta) - J^{\mathrm{fr}}_{k,t}(\theta)| &= \left|\frac{1}{n_k}\sum_i \big(w_{k,i}(\theta; \Lambda^t) - w^{t,\star}_{k,i}\big)\ell_{k,i}(\theta)\right| \\ &\le \frac{1}{2\alpha n_k}\sum_i |\ell_{k,i}(\theta) - \ell_{k,i}(\theta^t)|\ell_{k,i}(\theta) \\ &\le \frac{BL}{2\alpha}\|\theta - \theta^t\|_2. \end{aligned} \tag{9}$$

If $\|\theta - \theta^t\|_2 \le R_{k,t}$ throughout $\mathcal{T}_{k,t}$, then the objective mismatch is uniformly bounded by $BLR_{k,t}/(2\alpha)$. For a client taking $E$ local steps, write its iterates as $\theta^t_{k,0} = \theta^t, \ldots, \theta^t_{k,E}$. If every step has Euclidean length at most $s$, the triangle inequality gives

$$\|\theta^t_{k,e} - \theta^t\|_2 \le \sum_{j=0}^{e-1}\|\theta^t_{k,j+1} - \theta^t_{k,j}\|_2 \le es, \qquad e = 0, \ldots, E.$$

Thus, the mismatch at any local iterate is at most $BLEs/(2\alpha)$.

# D RESIDUAL ALIGNMENT AND PRICE RESPONSE

We derive the relation between smooth residuals and channel losses, then examine how a price update changes the sample weights. With fixed model $\theta$ throughout, all the channel means are empirical averages over the pooled training data, and empty label slices are omitted.

## D.1 LOSS DEVIATIONS AND SMOOTH RESIDUALS

Write $p_i = \sigma(z_i)$ as the output probability, where $z_i$ is a finite sample logit, $\sigma(z) = (1 + e^{-z})^{-1}$. The binary cross-entropy losses as functions of the logit are

$$b_1(z) = -\log \sigma(z), \qquad b_0(z) = -\log(1 - \sigma(z)).$$

For $g \in A_y$ and $m_y = |A_y|$, define the channel means and their label-wise references:

$$\bar{z}_{g,y} = \frac{1}{n_{g,y}} \sum_{i:s_i=g,y_i=y} z_i, \qquad \bar{z}_{A_y} = \frac{1}{m_y} \sum_{g \in A_y} \bar{z}_{g,y},$$
$$L_{g,y} = \frac{1}{n_{g,y}} \sum_{i:s_i=g,y_i=y} b_y(z_i), \quad \bar{L}_y = \frac{1}{m_y} \sum_{g \in A_y} L_{g,y}.$$

The references average groups equally regardless of channel size. The smooth rates are

$$\tau_g = \frac{1}{n_{g,1}} \sum_{i:s_i=g,y_i=1} p_i, \qquad \phi_g = \frac{1}{n_{g,0}} \sum_{i:s_i=g,y_i=0} p_i,$$

with residuals $\widetilde{c}_g^{+} = \bar{\tau}_{A_1} - \tau_g$ and $\widetilde{c}_g^{-} = \phi_g - \bar{\phi}_{A_0}$, where the bars also denote arithmetic means over active groups.

Assume the local logit concentration and non-saturation conditions: for some $\varepsilon_z > 0$ and $\kappa \in (0, 1/4]$, every active channel satisfies

$$\frac{1}{n_{g,y}} \sum_{i:s_i=g,y_i=y} (z_i - \bar{z}_{A_y})^2 \le \varepsilon_z^2, \qquad \sigma'(\bar{z}_{A_y}) \ge \kappa$$

Set $q_y = \sigma(\bar{z}_{A_y})$ and $\delta_{g,y} = \bar{z}_{g,y} - \bar{z}_{A_y}$. Then $\sum_{g \in A_y} \delta_{g,y} = 0$. For each $h \in \{\sigma, b_0, b_1\}$, the second derivative satisfies $|h''(z)| \le 1/4$. Taylor expansion at $\bar{z}_{A_y}$ therefore gives

$$\frac{1}{n_{g,y}} \sum_{i:s_i=g,y_i=y} h(z_i) = h(\bar{z}_{A_y}) + h'(\bar{z}_{A_y})\delta_{g,y} + R^h_{g,y}, \qquad |R^h_{g,y}| \le \frac{\varepsilon_z^2}{8}.$$

Subtracting the label-wise group mean cancels the constant term and the mean logit deviation. The resulting remainder is at most $\varepsilon_z^2/4$ in magnitude. Using $b_1'(z) = \sigma(z) - 1$, $b_0'(z) = \sigma(z)$, and $\sigma'(z) = \sigma(z)(1 - \sigma(z))$, we obtain

$$L_{g,1} - \bar{L}_1 = -(1 - q_1)\delta_{g,1} + R^L_{g,1}, \quad \widetilde{c}_g^{+} = -q_1(1 - q_1)\delta_{g,1} + R^c_{g,1},$$
$$L_{g,0} - \bar{L}_0 = q_0\delta_{g,0} + R^L_{g,0}, \qquad \widetilde{c}_g^{-} = q_0(1 - q_0)\delta_{g,0} + R^c_{g,0},$$

where $|R^L_{g,y}|, |R^c_{g,y}| \le \varepsilon_z^2/4$. Eliminating the logit deviations yields

$$L_{g,1} - \bar{L}_1 = \frac{\widetilde{c}_g^{+}}{q_1} + R_{g,1}, \qquad L_{g,0} - \bar{L}_0 = \frac{\widetilde{c}_g^{-}}{1 - q_0} + R_{g,0}. \tag{10}$$

Since $q_y(1 - q_y) \ge \kappa$, both $q_y$ and $1 - q_y$ are at least $\kappa$. Thus the remainders satisfy the uniform bound

$$|R_{g,y}| \le C_\kappa \varepsilon_z^2, \qquad C_\kappa = \frac{1 + \kappa^{-1}}{4}.$$

This proves the first-order alignment result. Both leading coefficients in (10) are positive, so a channel's loss deviation and smooth residual have the same sign whenever the leading term has magnitude greater than $C_\kappa \varepsilon_z^2$.

### D.2 Response of channel attention to prices

Hold the sample losses $\ell_i = \ell_i(\theta)$ fixed. By the allocation rule (8), the average attention in channel $(g, y)$ at price $\lambda$ is

$$Q_{g,y}(\lambda) = \frac{1}{2\alpha n_{g,y}} \sum_{i:s_i=g,y_i=y} (\ell_i - \lambda)_+, \qquad \alpha > 0.$$

Its total attention is $W_{g,y} = n_{g,y} Q_{g,y}(\lambda_{g,y})$. Let

$$\mathcal{I}_{g,y}(\lambda) = \{i : s_i = g,\ y_i = y,\ \ell_i > \lambda\}, \qquad a_{g,y}(\lambda) = \frac{|\mathcal{I}_{g,y}(\lambda)|}{n_{g,y}}$$

be the set and fraction of samples with positive weights. $(\ell_i - \lambda)_+$ has slope $-1$ below $\ell_i$ and zero above it. Integrating these slopes and averaging gives, for any price change $\delta\lambda$,

$$Q_{g,y}(\lambda + \delta\lambda) - Q_{g,y}(\lambda) = -\frac{1}{2\alpha} \int_{\lambda}^{\lambda+\delta\lambda} a_{g,y}(u)\, du. \tag{11}$$

Hence $Q_{g,y}$ is non-increasing and $1/(2\alpha)$-Lipschitz in its price, including at truncation thresholds. If the positive-weight set is unchanged throughout the interval, then

$$Q_{g,y}(\lambda + \delta\lambda) - Q_{g,y}(\lambda) = -a_{g,y}(\lambda) \frac{\delta\lambda}{2\alpha}.$$

When every weight remains positive, $a_{g,y} = 1$; when every weight remains zero, a price change within the interval has no effect.

For the round-$t$ controller update, take the fixed model to be $\theta^{t+1}$ and consider smooth feedback, $\beta_t = 0$. Write $s_1 = \gamma_t \eta_{\mathrm{tpr}} > 0$ and $s_0 = \gamma_t \eta_{\mathrm{fpr}} > 0$. The raw updates are

$$\delta\lambda_{g,1} = -s_1 \widetilde{c}_g^{+}, \qquad \delta\lambda_{g,0} = -s_0 \widetilde{c}_g^{-}.$$

A positive residual therefore lowers the price and cannot decrease the channel's average attention. On an unchanged positive-weight set, the changes are

$$\Delta Q_{g,1} = \frac{s_1 a_{g,1}(\lambda_{g,1})}{2\alpha} \widetilde{c}_g^{+}, \qquad \Delta Q_{g,0} = \frac{s_0 a_{g,0}(\lambda_{g,0})}{2\alpha} \widetilde{c}_g^{-}.$$

The increase is strict when the residual and the positive-weight fraction are both positive. These formulas also describe the projected update whenever the raw update is feasible. Indeed, each label slice of the residual sums to zero, so the common step size within that slice preserves price centering. Starting from a feasible controller, if all updated prices remain in $[-\lambda_{\max}, \lambda_{\max}]$, the projection leaves them unchanged. If clipping is required, the centering adjustment couples the coordinates, and the direction of an individual price change need not be preserved.

### D.3 Hard and smooth residuals

For hybrid feedback, the sign of a hard residual can differ from that of its smooth counterpart. The following bound gives a sufficient condition for agreement using the scores at the same model.

**Lemma 2** (Hard and smooth residual agreement)**.** *Let $\widehat{y}_i = \mathbb{I}\{p_i \geq 1/2\}$, and let $c_g^{+}$ and $c_g^{-}$ be the corresponding hard residuals. Define*

$$e_{g,y} = \frac{1}{n_{g,y}} \sum_{i:s_i=g,y_i=y} |p_i - \widehat{y}_i|, \qquad \bar{e}_y = \frac{1}{m_y} \sum_{g \in A_y} e_{g,y}.$$

*Then*

$$|c_g^{+} - \widetilde{c}_g^{+}| \leq e_{g,1} + \bar{e}_1, \qquad |c_g^{-} - \widetilde{c}_g^{-}| \leq e_{g,0} + \bar{e}_0.$$

*In either label slice, if the magnitude of the smooth residual exceeds the corresponding bound, the hard and smooth residuals have the same sign. Every mixture $r_g^{\pm} = \beta_t c_g^{\pm} + (1 - \beta_t)\widetilde{c}_g^{\pm}$, $\beta_t \in [0, 1]$, then has that sign as well.*

*Proof.* Let

$$d_{g,y} = \frac{1}{n_{g,y}} \sum_{i:s_i=g,y_i=y} (\widehat{y}_i - p_i), \qquad \bar{d}_y = \frac{1}{m_y} \sum_{g \in A_y} d_{g,y}.$$

The triangle inequality gives $|d_{g,y}| \le e_{g,y}$ and $|\bar{d}_y| \le \bar{e}_y$. From the definitions of the residuals,

$$c_g^+ - \widetilde{c}_g^+ = \bar{d}_1 - d_{g,1}, \qquad c_g^- - \widetilde{c}_g^- = d_{g,0} - \bar{d}_0.$$

Applying the triangle inequality proves both bounds. If the smooth residual has magnitude strictly larger than its error bound, the hard residual cannot be zero or have the opposite sign. A convex combination of two nonzero values with the same sign retains that sign. □

Under this condition, hybrid feedback gives the same raw price-update direction as smooth feedback; the truncation and projection conditions above still apply. These results concern the allocation at a fixed model. Changes in EO or EOD after local training depend on the resulting model update.

## E STRICT CONTROLLER EXPRESSIVITY

For EOD criterion, we compare label-specific channel prices with a controller that assigns the same price to both labels of each group. As in the main text, assume $A_0 = A_1 = A$. For $\lambda_{\max} \ge 0$, the feasible controller set is

$$\mathcal{C} = \left\{ \Lambda : \begin{array}{ll} |\lambda_{g,y}| \le \lambda_{\max}, & \\ \lambda_{g,y} = 0 & (g \notin A), \\ \sum_{g \in A} \lambda_{g,y} = 0 & (y \in \{0,1\}) \end{array} \right\},$$

and tying the two prices within each group gives

$$\mathcal{C}_{\text{coarse}} = \{\Lambda \in \mathcal{C} : \lambda_{g,0} = \lambda_{g,1} \text{ for every } g \in A\}.$$

Both sets contain the zero controller and are compact, so a linear objective attains its maximum on each. For any fixed direction $D$ supported on $A \times \{0,1\}$ and centered within each label slice, the inclusion $\mathcal{C}_{\text{coarse}} \subseteq \mathcal{C}$ gives

$$\max_{\Lambda \in \mathcal{C}} \langle \Lambda, D \rangle \ge \max_{\Lambda \in \mathcal{C}_{\text{coarse}}} \langle \Lambda, D \rangle.$$

To show that the inequality can be strict, suppose $|A| \ge 2$ and $\lambda_{\max} > 0$. Choose distinct groups $a, b \in A$, fix $d > 0$, and set

$$D_{a,1} = d, \qquad D_{b,1} = -d, \qquad D_{a,0} = -d, \qquad D_{b,0} = d,$$

with all other entries of $D$ equal to zero. The direction is supported on the active channels and sums to zero in each label slice. It also satisfies $D_{g,0} + D_{g,1} = 0$ for every group. Hence, for any tied controller, $\lambda_{g,0} = \lambda_{g,1} = \lambda_g$, we have

$$\langle \Lambda, D \rangle = \sum_{g \in A} \lambda_g (D_{g,0} + D_{g,1}) = 0.$$

For any $\Lambda \in \mathcal{C}$, the box constraints give

$$\langle \Lambda, D \rangle = d(\lambda_{a,1} - \lambda_{b,1} - \lambda_{a,0} + \lambda_{b,0}) \le 4d\lambda_{\max}.$$

This upper bound is attained by choosing

$$\lambda^\star_{a,1} = \lambda^\star_{b,0} = \lambda_{\max}, \qquad \lambda^\star_{b,1} = \lambda^\star_{a,0} = -\lambda_{\max},$$

with all other prices equal to zero. Each label slice sums to zero and every price satisfies the box constraint, so $\Lambda^\star \in \mathcal{C}$. Thus, for this direction,

$$\max_{\Lambda \in \mathcal{C}} \langle \Lambda, D \rangle = 4d\lambda_{\max} > 0 = \max_{\Lambda \in \mathcal{C}_{\text{coarse}}} \langle \Lambda, D \rangle.$$

The construction applies to any common active set with at least two groups, since additional groups receive zero direction entries and prices. If $|A| \le 1$ or $\lambda_{\max} = 0$, both feasible sets reduce to the zero controller, and strict inequality is impossible.

# F CONTROLLER BOUNDEDNESS AND ONLINE REGRET

Let $q$ count controller-active updates, and identify $\Lambda^q$ with its vector of $2G$ channel prices. The active sets $A_y$ are fixed throughout this section and may differ across labels. The feasible set $\mathcal{C}$ is the centered box from the main text: inactive coordinates are zero, prices sum to zero within each label slice, and $|\lambda_{g,y}| \le \lambda_{\max}$, where $\lambda_{\max} \ge 0$.

The projection separates across label slices, with inactive coordinates fixed at zero. The following lemma justifies the formula in the main text.

**Lemma 3** (Projection onto a centered box)**.** *Let $m \ge 1$, $L \ge 0$, $v \in \mathbb{R}^m$, and*

$$\mathcal{C}_m = \left\{ \lambda \in [-L, L]^m : \sum_{i=1}^{m} \lambda_i = 0 \right\}.$$

*The unique Euclidean projection of $v$ onto $\mathcal{C}_m$ is $\lambda_i^\star = \mathrm{clip}(v_i - \zeta, -L, L)$, where $\zeta$ is any zero of*

$$\psi(\zeta) = \sum_{i=1}^{m} \mathrm{clip}(v_i - \zeta, -L, L).$$

*Such a zero exists in $[\min_i v_i - L, \max_i v_i + L]$. The shift $\zeta$ need not be unique, but all such zeros give the same projected vector.*

*Proof.* The feasible set is nonempty, compact, and convex, so the strictly convex objective $\frac{1}{2}\|\lambda - v\|_2^2$ has a unique minimizer. The function $\psi$ is continuous and non-increasing. Writing $a = \min_i v_i - L$ and $b = \max_i v_i + L$, we have $\psi(a) = mL$ and $\psi(b) = -mL$, which ensures a zero in $[a, b]$.

For a scalar multiplier $\zeta$, minimizing

$$\frac{1}{2}\|\lambda - v\|_2^2 + \zeta \sum_{i=1}^{m} \lambda_i$$

over $[-L, L]^m$ separates across coordinates and gives $\lambda_i = \mathrm{clip}(v_i - \zeta, -L, L)$. At a zero of $\psi$, this vector is feasible. The multiplier term vanishes for every feasible vector, so this minimizer also minimizes the original projection objective over $\mathcal{C}_m$. Uniqueness of the projection implies that every zero yields the same vector. □

**Computing the shift.** Apply the lemma to each active label slice with $m = |A_y|$, $L = \lambda_{\max}$, and $v = (v_{g,y}^{t+1})_{g \in A_y}$. Skip empty slices; if $m = 1$ or $L = 0$, set the slice to zero. Otherwise, choose a tolerance $\varepsilon > 0$ and initialize

$$a = \min_i v_i - L, \qquad b = \max_i v_i + L.$$

While $b - a > 2\varepsilon/m$, let $c = (a + b)/2$. If $\psi(c) = 0$, return $\mathrm{clip}(v - c, -L, L)$. If $\psi(c) > 0$, set $a = c$; otherwise, set $b = c$. Upon termination, set $\widehat{\zeta} = (a+b)/2$ and return $\widehat{\lambda} = \mathrm{clip}(v - \widehat{\zeta}, -L, L)$, with clipping applied coordinate wise.

Each retained interval contains a zero of $\psi$. Since clipping is 1-Lipschitz and $\psi$ is $m$-Lipschitz, the stopping rule gives

$$\|\widehat{\lambda} - \lambda^\star\|_\infty \le \frac{\varepsilon}{m}, \qquad \left| \sum_i \widehat{\lambda}_i \right| \le \varepsilon.$$

These bounds quantify the numerical accuracy of finite-iteration bisection. We now analyze the controller updates using the exact Euclidean projection $\Pi_{\mathcal{C}}$, obtained by applying Lemma 3 to each active label slice and setting inactive coordinates to zero.

Index controller-active updates by $q = 0, 1, \ldots$, and let $t_q$ be the communication round containing update $q$. The controller $\Lambda^q$ is the state immediately before this update, and the feedback is evaluated at $\theta^{t_q+1}$. Define the round-specific mixed residuals by

$$r_g^{\pm,q} = \beta_{t_q} c_g^{\pm}(\theta^{t_q+1}) + (1 - \beta_{t_q})\bar{c}_g^{\pm}(\theta^{t_q+1}), \qquad \beta_{t_q} \in [0, 1].$$

For fixed finite label step sizes, the update direction in EOD mode is

$$\widetilde{d}^q_{g,1} = -\eta_{\mathrm{tpr}} r_g^{+,q}, \quad g \in A_1,$$
$$\widetilde{d}^q_{g,0} = -\eta_{\mathrm{fpr}} r_g^{-,q}, \quad g \in A_0,$$

with inactive coordinates set to zero. Starting from $\Lambda^0 \in \mathcal{C}$, for example at zero, the projected controller update is

$$\Lambda^{q+1} = \Pi_{\mathcal{C}}\big(\Lambda^q + \gamma_q \widetilde{d}^q\big), \qquad \gamma_q = (1+\rho q)^{-1/2}, \qquad \rho > 0.$$

Hard prediction rates and mean probability scores lie in $[0,1]$. Their centered residuals therefore lie in $[-1,1]$, as do the convex mixtures $r_g^{\pm,q}$. Consequently,

$$\|\widetilde{d}^q\|_2^2 \le |A_1|\eta_{\mathrm{tpr}}^2 + |A_0|\eta_{\mathrm{fpr}}^2 =: M_0^2. \tag{12}$$

For EO criterion, set $\eta_{\mathrm{fpr}} = 0$ and initialize all negative-label prices at zero. The negative-label update direction is then identically zero. Since the projection separates across label slices and the zero vector is feasible, the negative-label prices remain zero after every update. Consequently, the direction bound reduces to

$$M_0^2 = |A_1|\eta_{\mathrm{tpr}}^2.$$

The online regret theorem below therefore applies to EO with this value of $M_0$, using the same feasible set $\mathcal{C}$. Thus, bounded update directions follow directly from the feedback construction.

Feasible initialization and projection ensure that $\Lambda^q \in \mathcal{C}$ for every $q \ge 0$. The coordinate bounds therefore give

$$\|\Lambda^q\|_2 \le \lambda_{\max}\sqrt{|A_0| + |A_1|} \le \lambda_{\max}\sqrt{2G}.$$

**Theorem 1** (Online regret)**.** *Let $(\widetilde{d}^q)_{q\ge 0}$ be any sequence satisfying $\|\widetilde{d}^q\|_2 \le M_0$ for some finite $M_0 \ge 0$, and consider the projected update above with $\Lambda^0 \in \mathcal{C}$. Let $\Delta = \mathrm{diam}(\mathcal{C})$. For every fixed comparator $\Lambda \in \mathcal{C}$ and integer $Q \ge 1$,*

$$\sum_{q=0}^{Q-1} \langle \Lambda - \Lambda^q, \widetilde{d}^q \rangle \le \frac{\Delta^2}{2\gamma_{Q-1}} + \frac{M_0^2}{2}\sum_{q=0}^{Q-1}\gamma_q. \tag{13}$$

*For fixed $\Delta$, $M_0$, and $\rho > 0$, the right-hand side is $O(\sqrt{Q})$.*

*Proof.* Set $a_q = \|\Lambda - \Lambda^q\|_2^2$. The nonempty set $\mathcal{C}$ is closed and convex, being the intersection of a box and linear equality constraints. Since $\Pi_{\mathcal{C}}(\Lambda) = \Lambda$, non-expansiveness of Euclidean projection gives

$$\begin{aligned} a_{q+1} &\le \|\Lambda - (\Lambda^q + \gamma_q \widetilde{d}^q)\|_2^2 \\ &= a_q - 2\gamma_q \langle \Lambda - \Lambda^q, \widetilde{d}^q \rangle + \gamma_q^2 \|\widetilde{d}^q\|_2^2. \end{aligned}$$

Using $\|\widetilde{d}^q\|_2 \le M_0$, we obtain

$$\langle \Lambda - \Lambda^q, \widetilde{d}^q \rangle \le \frac{a_q - a_{q+1}}{2\gamma_q} + \frac{\gamma_q M_0^2}{2}.$$

Both $\Lambda$ and $\Lambda^q$ belong to $\mathcal{C}$, so $0 \le a_q \le \Delta^2$. Summing the distance terms with their varying denominators gives

$$\begin{aligned} \sum_{q=0}^{Q-1} \frac{a_q - a_{q+1}}{2\gamma_q} &= \frac{a_0}{2\gamma_0} + \sum_{q=1}^{Q-1} a_q \left(\frac{1}{2\gamma_q} - \frac{1}{2\gamma_{q-1}}\right) - \frac{a_Q}{2\gamma_{Q-1}} \\ &\le \Delta^2 \left[\frac{1}{2\gamma_0} + \sum_{q=1}^{Q-1}\left(\frac{1}{2\gamma_q} - \frac{1}{2\gamma_{q-1}}\right)\right] \\ &= \frac{\Delta^2}{2\gamma_{Q-1}}. \end{aligned}$$

The inequality uses that $\gamma_q$ is non-increasing, so every coefficient in the middle sum is nonnegative, while the final term is non-positive. Summing the one-step bound proves (13). For the specified step sizes, $\gamma_{Q-1}^{-1} = \sqrt{1+\rho(Q-1)}$. Since $(1+\rho u)^{-1/2}$ is decreasing,

$$\sum_{q=0}^{Q-1} \gamma_q \le 1 + \int_0^{Q-1} (1+\rho u)^{-1/2}\, du$$
$$= 1 + \frac{2}{\rho}\big(\sqrt{1+\rho(Q-1)} - 1\big).$$

Therefore, both terms on the right-hand side of (13) are $O(\sqrt{Q})$ for fixed $\rho > 0$. □

The regret compares $\Lambda^q$ with a fixed feasible controller under the linear rewards $\langle \Lambda, \widetilde{d}^q \rangle$. For the actual feedback defined above, these rewards use the negative fairness residuals scaled by the label step sizes, with every comparator evaluated on the same realized feedback sequence.

## G A CONDITIONAL BOUND ON AVERAGED SMOOTH RESIDUALS

The regret bound in Appendix F controls linear controller rewards. To obtain a bound on averaged smooth residuals, we impose an additional comparison condition on the model update. The results are first presented for EOD criterion with positive step sizes, then we give the EO specialization at the end of this section.

As in Appendix F, let $q$ count controller-active updates. Write $x_q$ for the post-aggregation model used to compute the feedback at update $q$. If that update occurs in communication round $t_q$, then $x_q = \theta^{t_q+1}$. For each active channel, write

$$\widetilde{c}_{g,1}(\theta) = \widetilde{c}_g^{+}(\theta), \qquad \widetilde{c}_{g,0}(\theta) = \widetilde{c}_g^{-}(\theta),$$

and, for fixed positive label step sizes, define

$$s_{g,y}(\theta) = \eta_y \widetilde{c}_{g,y}(\theta), \qquad \eta_1 = \eta_{\mathrm{tpr}} > 0, \quad \eta_0 = \eta_{\mathrm{fpr}} > 0.$$

Inactive entries of $s$ are zero, and empty label slices are omitted. Since the step size is shared within each label, $s(\theta)$ is centered in each active slice.

Use the fixed controller set $\mathcal{C}$ from Appendix F, with $\lambda_{\max} > 0$ and $\Lambda^0 \in \mathcal{C}$. In pure smooth mode, the projected update has direction $-s(x_q)$. Set $\Gamma^q = -\Lambda^q$. The set is centrally symmetric, $\mathcal{C} = -\mathcal{C}$; changing variables in the projection problem gives $\Pi_{\mathcal{C}}(-v) = -\Pi_{\mathcal{C}}(v)$. Hence

$$\Gamma^{q+1} = -\Pi_{\mathcal{C}}\big(\Lambda^q - \gamma_q s(x_q)\big) = \Pi_{\mathcal{C}}\big(\Gamma^q + \gamma_q s(x_q)\big), \qquad \Gamma^0 \in \mathcal{C},$$

where $\gamma_q = (1+\rho q)^{-1/2}$ and $\rho > 0$.

Let $F$ be the pooled empirical risk and define the reference Lagrangian

$$\mathcal{L}_{\mathrm{oracle}}(\theta, \Gamma) = F(\theta) + \langle \Gamma, s(\theta) \rangle.$$

**Assumption 1** (Reference comparison)**.** There is a fixed model $\theta^\dagger$ with finite risk and $s(\theta^\dagger) = 0$, and nonnegative errors $\varepsilon_q$, such that

$$\mathcal{L}_{\mathrm{oracle}}(x_q, \Gamma^q) \le \mathcal{L}_{\mathrm{oracle}}(\theta^\dagger, \Gamma^q) + \varepsilon_q = F(\theta^\dagger) + \varepsilon_q$$

at every controller update.

This is an additional condition on the models $x_q$. GloReFair's frozen-weight local objective does not establish the comparison with $\mathcal{L}_{\mathrm{oracle}}$.

Let $\Delta = \mathrm{diam}(\mathcal{C})$ and choose $M_0 > 0$ such that $\|s(x_q)\|_2 \le M_0$ for every $q$. Such a bound exists for fixed step sizes because every smooth residual lies in $[-1, 1]$. The regret bound in Theorem 1 therefore applies to $\Gamma^q$ with feedback $s(x_q)$. For any vector $v$, denote the support function of $\mathcal{C}$ by

$$\sigma_{\mathcal{C}}(v) = \max_{\Gamma \in \mathcal{C}} \langle \Gamma, v \rangle.$$

**Proposition 4** (Averaged residual bound)**.** *Under Assumption 1 and the setup above, fix an integer* $Q \geq 1$ *and define*

$$\bar{s}_Q = \frac{1}{Q}\sum_{q=0}^{Q-1} s(x_q), \qquad \bar{\varepsilon}_Q = \frac{1}{Q}\sum_{q=0}^{Q-1} \varepsilon_q,$$

$$B_Q = \frac{\Delta^2}{2\gamma_{Q-1}} + \frac{M_0^2}{2}\sum_{q=0}^{Q-1}\gamma_q, \qquad \delta_Q^+ = \left[F(\theta^\dagger) - \frac{1}{Q}\sum_{q=0}^{Q-1} F(x_q)\right]_+.$$

*Then*

$$\sigma_{\mathcal{C}}(\bar{s}_Q) \leq \frac{B_Q}{Q} + \bar{\varepsilon}_Q + \delta_Q^+. \tag{14}$$

*For each nonempty label slice,*

$$\max_{g\in A_y}\left|\frac{1}{Q}\sum_{q=0}^{Q-1}\widetilde{c}_{g,y}(x_q)\right| \leq \frac{1}{\lambda_{\max}\eta_y}\left(\frac{B_Q}{Q} + \bar{\varepsilon}_Q + \delta_Q^+\right). \tag{15}$$

*Proof.* Assumption 1 gives

$$F(x_q) + \langle \Gamma^q, s(x_q)\rangle \leq F(\theta^\dagger) + \varepsilon_q.$$

For every fixed $\Gamma \in \mathcal{C}$, Theorem 1 gives

$$\sum_{q=0}^{Q-1}\langle \Gamma - \Gamma^q, s(x_q)\rangle \leq B_Q.$$

Summing the model comparison over $q$, adding these inequalities, and dividing by $Q$ yields

$$\langle \Gamma, \bar{s}_Q\rangle \leq \frac{B_Q}{Q} + \bar{\varepsilon}_Q + F(\theta^\dagger) - \frac{1}{Q}\sum_{q=0}^{Q-1} F(x_q).$$

Maximizing over $\Gamma \in \mathcal{C}$ and bounding the last risk difference by its positive part proves (14).

For the coordinate bound, fix a nonempty label slice and write $z_g = \bar{s}_{Q,g,y}$. These coordinates sum to zero. If all are zero, the bound is immediate; this includes every slice with a single active group. Otherwise, choose $g_+$ and $g_-$ attaining the maximum and minimum of $z_g$. Centering implies $z_{g_+} > 0 > z_{g_-}$, so these indices are distinct. Define a controller $\widehat{\Gamma}$ with value $+\lambda_{\max}$ at $(g_+, y)$, value $-\lambda_{\max}$ at $(g_-, y)$, and zero elsewhere. This controller belongs to $\mathcal{C}$, and therefore

$$\begin{aligned}\sigma_{\mathcal{C}}(\bar{s}_Q) &\geq \langle \widehat{\Gamma}, \bar{s}_Q\rangle \\ &= \lambda_{\max}(z_{g_+} - z_{g_-}) \geq \lambda_{\max}\max_{g\in A_y}|z_g|.\end{aligned}$$

Since $z_g = \eta_y \frac{1}{Q}\sum_{q=0}^{Q-1}\widetilde{c}_{g,y}(x_q)$, we obtain (15) according to $\lambda_{\max}\eta_y > 0$ and (14). □

**Interpretation and scope.** Define the score ensemble over the $Q$ controller-active updates by

$$\bar{f}_Q(\xi) = \frac{1}{Q}\sum_{q=0}^{Q-1} f_{x_q}(\xi).$$

For a predictor $f$ with scores in $[0,1]$, write $\widetilde{c}_{g,y}[f]$ for the smooth residual computed from its scores using the same empirical channel averages and group-centering rule. Since these operations are linear on the fixed pooled dataset,

$$\widetilde{c}_{g,y}[\bar{f}_Q] = \frac{1}{Q}\sum_{q=0}^{Q-1}\widetilde{c}_{g,y}(x_q).$$

Thus, (15) bounds the maximum absolute empirical smooth residual of $\bar{f}_Q$ within each label slice. The ensemble is defined at the score level and need not be representable by a single parameter vector in the original model family.

Because signed residuals may cancel across updates, this result does not control the average per-update residual magnitude, or the residuals of the last iterate or a model obtained by averaging parameters. Hard EO or EOD of the threshold-based ensemble additionally depends on the hard-smooth discrepancy; Lemma 2 provides a bound when evaluated using the ensemble scores and their threshold-based predictions.

The regret contribution $B_Q/Q$ is $O(Q^{-1/2})$. The other terms are the average comparison error and the positive part of the reference risk minus the average risk of $x_q$. A vanishing residual bound therefore additionally requires $\bar{\varepsilon}_Q \to 0$ and $\delta_Q^+ \to 0$.

**EO specialization.** Assume $A_1 \neq \varnothing$, $\eta_{\mathrm{tpr}} > 0$, and $\lambda_{\max} > 0$. For EO criterion, retain the same feasible set $\mathcal{C}$ and define

$$s_{g,1}^{\mathrm{EO}}(\theta) = \eta_{\mathrm{tpr}} \widetilde{c}_g^+(\theta), \qquad g \in A_1, \qquad s_{g,0}^{\mathrm{EO}}(\theta) = 0,$$

with inactive positive-label entries also set to zero. Initialize the negative-label prices at zero. Under pure smooth feedback, suppose Assumption 1 holds for the EO model sequence with $s$ replaced throughout by $s^{\mathrm{EO}}$. The reference model is therefore required to have finite pooled risk and zero positive-label smooth residuals, while $F$ continues to include samples of both labels. The preceding proof then gives

$$\max_{g \in A_1} \left| \frac{1}{Q} \sum_{q=0}^{Q-1} \widetilde{c}_g^+(x_q) \right| \leq \frac{B_Q/Q + \bar{\varepsilon}_Q + \delta_Q^+}{\lambda_{\max} \eta_{\mathrm{tpr}}}.$$

Here $B_Q$, $\bar{\varepsilon}_Q$ and $\delta_Q^+$ are evaluated using the EO trajectory and its reference comparison, with $M_0^2 = |A_1| \eta_{\mathrm{tpr}}^2$ and $\Delta = \mathrm{diam}(\mathcal{C})$. By the score-ensemble identity above, this bounds the maximum absolute positive-label smooth residual of the EO score ensemble.

# H Experimental Details and Additional Results

## H.1 Experimental Protocol

We use Adult, CreditCard, and Dutch, with sex as the sensitive attribute. Unless otherwise stated, training uses five clients, full participation, 100 communication rounds, 10 local epochs per round, and a batch size of 128. We consider group, label, and joint Dirichlet partitions with concentration parameter $H \in \{1.0, 0.8, 0.6, 0.4, 0.2\}$; smaller $H$ produces stronger heterogeneity. Results are means over seeds $\{42, 123, 456\}$, with standard deviations in parentheses.

**Model, optimization, and prediction.** We implement GloReFair with a logistic regression (LR) classifier using a sigmoid output and binary cross-entropy loss. Local training is performed using mini-batch stochastic gradient descent (SGD), with the learning rate selected through the hyperparameter search described below. Binary predictions are obtained using a fixed threshold of 0.5:

$$\widehat{y}_i = \mathbf{1}\{f_\theta(x_i) \geq 0.5\}.$$

These predictions are used to compute accuracy, EO and EOD gaps, and the hard residuals used by the controller. Smooth residuals are computed directly from the predicted probabilities.

**Hyperparameter search.** For each dataset, method, and fairness criterion, we reserve a random 20% of the training data for validation and partition the remaining 80% IID across five clients. Each method receives 25 candidate configurations generated by a deterministic shifted-Halton design, including its default configuration and a configuration with fairness control disabled where available. All methods search the local learning rate log-uniformly over $[10^{-3}, 3 \times 10^{-2}]$. The method-specific search spaces are:

| Method | Search space |
|---|---|
| FairFed | $\beta \in \{0, 0.01, 0.05, 0.1, 1, 2, 5\}$ |
| FedFB | $\eta_\lambda \in \{0, 0.001, 0.01, 0.05, 0.1, 0.5\}$ |
| BGL | $B \in \{1, 5, 10, 20, 50, 100, 200\}$; tied thresholds $\zeta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ |
| Function Tracking | $\lambda_{\text{fair}} \in \{0, 0.01, 0.1, 1, 10\}$ |
| FedFACT, LoGoFair | global-local tolerance in $\{0, 0.005, 0.01, 0.02, 0.05\}$ |
| GloReFair | $\alpha \in [0.05, 1]$, $\eta_{\text{tpr}} \in [0.01, 0.2]$, $\lambda_{\max} \in \{0.5, 1, 2, 5\}$; $\eta_{\text{fpr}}/\eta_{\text{tpr}} \in \{0.1, 0.25, 0.5, 1, 2\}$ for EOD |

For GloReFair, we fix $\mu = 1$. Under EO, we set $\eta_{\text{fpr}} = 0$ and omit $\eta_{\text{fpr}}/\eta_{\text{tpr}}$ from the search.

**Model selection.** We discard failed or non-finite runs. To exclude degenerate classifiers, a candidate must have a validation predicted-positive rate in $[0.01, 0.99]$, sensitivity and specificity of at least 0.05, and balanced accuracy of at least 0.50. We select FedAvg by validation accuracy, breaking ties by balanced accuracy and then area under the precision–recall curve (AUPRC). For fairness-aware methods, we minimize the validation gap for the target criterion, breaking ties by accuracy, AUPRC, and balanced accuracy, in that order. Any remaining ties are resolved by candidate order. We reuse the selected configuration across partition families and heterogeneity levels; test data are not used for tuning or selection.

### H.2 Complete Comparison Results

Tables 3–5 report EOD results for group, label, and joint heterogeneity, respectively. Tables 6–8 report the corresponding EO results. Higher accuracy and lower fairness gaps are better. Red and green cells mark the best and second-best reported means in each column for each dataset. A dash indicates that the runs failed. And *Tracking* denotes Function Tracking method.

Table 3: The performance of different methods under the EOD criterion across datasets with varying group-shift severity.

| Method | Adult | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy (↑) | | | | | EOD Gap (↓) | | | | |
| | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ |
| FedAvg | 0.840 (0.000) | 0.840 (0.001) | 0.840 (0.001) | 0.841 (0.000) | 0.841 (0.001) | 0.125 (0.004) | 0.122 (0.004) | 0.117 (0.012) | 0.115 (0.011) | 0.122 (0.002) |
| FairFed | 0.838 (0.000) | 0.840 (0.001) | 0.839 (0.001) | 0.839 (0.001) | 0.840 (0.001) | 0.127 (0.007) | 0.127 (0.007) | 0.119 (0.012) | 0.118 (0.019) | 0.122 (0.005) |
| FedFB | 0.837 (0.001) | 0.838 (0.000) | 0.838 (0.001) | 0.839 (0.001) | 0.839 (0.000) | 0.052 (0.001) | 0.051 (0.001) | 0.050 (0.002) | 0.049 (0.003) | 0.052 (0.001) |
| BGL | 0.798 (0.001) | 0.799 (0.001) | 0.799 (0.000) | 0.799 (0.000) | 0.799 (0.001) | 0.075 (0.003) | 0.071 (0.001) | 0.076 (0.003) | 0.071 (0.004) | 0.072 (0.004) |
| Tracking | 0.834 (0.000) | 0.835 (0.001) | 0.834 (0.001) | 0.836 (0.000) | 0.835 (0.001) | 0.047 (0.004) | 0.045 (0.003) | 0.042 (0.002) | 0.046 (0.002) | 0.043 (0.002) |
| LoGoFair | 0.822 (0.001) | 0.815 (0.005) | 0.816 (0.005) | 0.714 (0.143) | 0.715 (0.050) | 0.037 (0.003) | 0.119 (0.042) | 0.080 (0.049) | 0.411 (0.407) | 0.463 (0.235) |
| FedFACT-in | 0.709 (0.011) | 0.699 (0.014) | 0.691 (0.018) | 0.686 (0.020) | – | 0.022 (0.023) | 0.017 (0.014) | 0.065 (0.003) | 0.062 (0.025) | – |
| FedFACT-post | 0.821 (0.001) | 0.823 (0.000) | 0.822 (0.001) | 0.823 (0.001) | – | 0.040 (0.002) | 0.043 (0.003) | 0.051 (0.014) | 0.045 (0.013) | – |
| GloReFair | 0.833 (0.001) | 0.833 (0.000) | 0.833 (0.001) | 0.833 (0.000) | 0.833 (0.000) | 0.042 (0.005) | 0.043 (0.003) | 0.041 (0.002) | 0.039 (0.003) | 0.041 (0.001) |
| | CreditCard | | | | | | | | | |
| | Accuracy (↑) | | | | | EOD Gap (↓) | | | | |
| | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ |
| FedAvg | 0.807 (0.001) | 0.805 (0.000) | 0.806 (0.000) | 0.806 (0.000) | 0.807 (0.001) | 0.063 (0.015) | 0.053 (0.014) | 0.056 (0.009) | 0.055 (0.011) | 0.056 (0.014) |
| FairFed | 0.805 (0.001) | 0.804 (0.000) | 0.805 (0.000) | 0.805 (0.000) | 0.805 (0.000) | 0.065 (0.011) | 0.056 (0.012) | 0.053 (0.008) | 0.056 (0.013) | 0.058 (0.012) |
| FedFB | 0.804 (0.001) | 0.803 (0.000) | 0.804 (0.000) | 0.804 (0.001) | 0.804 (0.001) | 0.065 (0.013) | 0.057 (0.009) | 0.058 (0.010) | 0.058 (0.011) | 0.061 (0.012) |
| BGL | 0.748 (0.001) | 0.747 (0.002) | 0.743 (0.004) | 0.746 (0.002) | 0.748 (0.007) | 0.010 (0.004) | 0.007 (0.003) | 0.015 (0.001) | 0.012 (0.002) | 0.012 (0.006) |
| Tracking | 0.804 (0.001) | 0.804 (0.000) | 0.805 (0.000) | 0.805 (0.001) | 0.805 (0.001) | 0.043 (0.005) | 0.041 (0.006) | 0.037 (0.005) | 0.037 (0.007) | 0.041 (0.007) |
| LoGoFair | 0.801 (0.001) | 0.796 (0.004) | 0.683 (0.165) | 0.723 (0.088) | 0.763 (0.038) | 0.037 (0.005) | 0.058 (0.039) | 0.219 (0.299) | 0.294 (0.281) | 0.144 (0.130) |
| FedFACT-in | 0.662 (0.022) | 0.670 (0.025) | 0.648 (0.016) | – | – | 0.066 (0.045) | 0.045 (0.027) | 0.061 (0.029) | – | – |
| FedFACT-post | 0.802 (0.001) | 0.802 (0.000) | 0.803 (0.001) | – | – | 0.061 (0.012) | 0.051 (0.007) | 0.056 (0.005) | – | – |
| GloReFair | 0.804 (0.001) | 0.804 (0.000) | 0.803 (0.000) | 0.803 (0.001) | 0.803 (0.000) | 0.032 (0.004) | 0.029 (0.001) | 0.033 (0.004) | 0.033 (0.000) | 0.033 (0.003) |
| | Dutch | | | | | | | | | |
| | Accuracy (↑) | | | | | EOD Gap (↓) | | | | |
| | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ |
| FedAvg | 0.834 (0.001) | 0.835 (0.001) | 0.834 (0.002) | 0.835 (0.002) | 0.835 (0.000) | 0.190 (0.002) | 0.185 (0.004) | 0.182 (0.007) | 0.183 (0.010) | 0.185 (0.006) |
| FairFed | 0.834 (0.002) | 0.836 (0.001) | 0.835 (0.001) | 0.835 (0.001) | 0.835 (0.000) | 0.190 (0.005) | 0.184 (0.003) | 0.195 (0.010) | 0.190 (0.002) | 0.191 (0.005) |
| FedFB | 0.821 (0.001) | 0.822 (0.001) | 0.821 (0.001) | 0.821 (0.001) | 0.822 (0.001) | 0.064 (0.007) | 0.057 (0.003) | 0.061 (0.003) | 0.057 (0.008) | 0.065 (0.006) |
| BGL | 0.825 (0.000) | 0.825 (0.000) | 0.826 (0.000) | 0.825 (0.001) | 0.825 (0.000) | 0.070 (0.003) | 0.073 (0.002) | 0.076 (0.003) | 0.078 (0.004) | 0.081 (0.004) |
| Tracking | 0.828 (0.001) | 0.829 (0.001) | 0.829 (0.001) | 0.828 (0.001) | 0.827 (0.002) | 0.070 (0.007) | 0.070 (0.003) | 0.067 (0.008) | 0.067 (0.011) | 0.066 (0.005) |
| LoGoFair | 0.809 (0.021) | 0.784 (0.026) | 0.757 (0.038) | 0.759 (0.064) | 0.699 (0.100) | 0.084 (0.043) | 0.185 (0.175) | 0.184 (0.146) | 0.209 (0.189) | 0.331 (0.178) |
| FedFACT-in | 0.628 (0.013) | 0.600 (0.014) | – | – | – | 0.022 (0.009) | 0.052 (0.047) | – | – | – |
| FedFACT-post | 0.827 (0.000) | 0.827 (0.001) | – | – | – | 0.056 (0.002) | 0.054 (0.004) | – | – | – |
| GloReFair | 0.825 (0.000) | 0.826 (0.000) | 0.826 (0.000) | 0.826 (0.001) | 0.825 (0.001) | 0.060 (0.003) | 0.059 (0.003) | 0.057 (0.001) | 0.066 (0.005) | 0.062 (0.007) |

Table 4: The performance of different methods under the EOD criterion across datasets with varying label-shift severity.

| Method | Adult | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy (↑) | | | | | EOD Gap (↓) | | | | |
| | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ |
| FedAvg | 0.834 (0.007) | 0.833 (0.003) | 0.810 (0.035) | 0.803 (0.036) | 0.790 (0.020) | 0.090 (0.034) | 0.085 (0.009) | 0.052 (0.033) | 0.040 (0.029) | 0.031 (0.008) |
| FairFed | 0.832 (0.006) | 0.832 (0.002) | 0.814 (0.034) | 0.801 (0.035) | 0.778 (0.024) | 0.121 (0.023) | 0.074 (0.010) | 0.102 (0.067) | 0.040 (0.041) | 0.101 (0.132) |
| FedFB | 0.832 (0.007) | 0.830 (0.003) | 0.808 (0.035) | 0.802 (0.035) | 0.789 (0.018) | 0.049 (0.008) | 0.033 (0.005) | 0.029 (0.007) | 0.032 (0.025) | 0.107 (0.010) |
| BGL | 0.799 (0.001) | 0.799 (0.001) | 0.799 (0.006) | 0.797 (0.005) | 0.803 (0.003) | 0.090 (0.011) | 0.075 (0.008) | 0.085 (0.004) | 0.095 (0.017) | 0.089 (0.011) |
| Tracking | 0.829 (0.004) | 0.829 (0.002) | 0.807 (0.036) | 0.802 (0.035) | 0.800 (0.015) | 0.032 (0.002) | 0.035 (0.004) | 0.039 (0.005) | 0.023 (0.014) | 0.075 (0.053) |
| LoGoFair | 0.729 (0.049) | 0.731 (0.037) | 0.678 (0.077) | 0.517 (0.095) | 0.718 (0.199) | 0.062 (0.011) | 0.032 (0.005) | 0.070 (0.042) | 0.252 (0.212) | 0.333 (0.260) |
| FedFACT-in | 0.666 (0.037) | 0.678 (0.016) | 0.617 (0.031) | 0.587 (0.113) | – | 0.014 (0.005) | 0.028 (0.017) | 0.009 (0.004) | 0.035 (0.021) | – |
| FedFACT-post | 0.814 (0.002) | 0.814 (0.007) | 0.799 (0.033) | 0.794 (0.031) | – | 0.032 (0.007) | 0.038 (0.014) | 0.025 (0.017) | 0.037 (0.033) | – |
| GloReFair | 0.827 (0.003) | 0.828 (0.002) | 0.819 (0.013) | 0.812 (0.020) | 0.810 (0.009) | 0.044 (0.004) | 0.040 (0.009) | 0.055 (0.027) | 0.067 (0.034) | 0.079 (0.011) |
| | CreditCard | | | | | | | | | |
| | Accuracy (↑) | | | | | EOD Gap (↓) | | | | |
| | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ |
| FedAvg | 0.800 (0.004) | 0.795 (0.011) | 0.793 (0.010) | 0.788 (0.007) | 0.788 (0.007) | 0.049 (0.007) | 0.036 (0.023) | 0.041 (0.030) | 0.046 (0.032) | 0.044 (0.030) |
| FairFed | 0.799 (0.003) | 0.794 (0.011) | 0.793 (0.010) | 0.788 (0.007) | 0.789 (0.007) | 0.048 (0.008) | 0.041 (0.023) | 0.041 (0.028) | 0.042 (0.030) | 0.046 (0.032) |
| FedFB | 0.798 (0.003) | 0.794 (0.010) | 0.792 (0.010) | 0.788 (0.007) | 0.789 (0.007) | 0.050 (0.008) | 0.042 (0.022) | 0.041 (0.027) | 0.044 (0.030) | 0.045 (0.031) |
| BGL | 0.741 (0.010) | 0.756 (0.033) | 0.762 (0.031) | 0.777 (0.026) | 0.733 (0.028) | 0.013 (0.005) | 0.017 (0.003) | 0.018 (0.003) | 0.011 (0.008) | 0.024 (0.020) |
| Tracking | 0.799 (0.002) | 0.794 (0.010) | 0.793 (0.010) | 0.790 (0.008) | 0.791 (0.009) | 0.028 (0.005) | 0.025 (0.014) | 0.026 (0.016) | 0.028 (0.020) | 0.034 (0.022) |
| LoGoFair | 0.731 (0.066) | 0.721 (0.027) | 0.696 (0.102) | 0.730 (0.103) | 0.457 (0.156) | 0.011 (0.002) | 0.295 (0.188) | 0.022 (0.008) | 0.019 (0.016) | 0.143 (0.170) |
| FedFACT-in | 0.618 (0.035) | 0.566 (0.079) | 0.612 (0.102) | – | – | 0.019 (0.007) | 0.015 (0.002) | 0.014 (0.006) | – | – |
| FedFACT-post | 0.795 (0.002) | 0.792 (0.008) | 0.792 (0.009) | – | – | 0.051 (0.005) | 0.039 (0.023) | 0.040 (0.022) | – | – |
| GloReFair | 0.798 (0.001) | 0.795 (0.008) | 0.794 (0.008) | 0.791 (0.006) | 0.792 (0.006) | 0.034 (0.001) | 0.033 (0.001) | 0.034 (0.002) | 0.031 (0.003) | 0.031 (0.002) |
| | Dutch | | | | | | | | | |
| | Accuracy (↑) | | | | | EOD Gap (↓) | | | | |
| | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ |
| FedAvg | 0.831 (0.003) | 0.834 (0.001) | 0.836 (0.001) | 0.832 (0.006) | 0.831 (0.007) | 0.168 (0.028) | 0.190 (0.013) | 0.204 (0.010) | 0.181 (0.067) | 0.179 (0.087) |
| FairFed | 0.828 (0.005) | 0.829 (0.005) | 0.835 (0.002) | 0.824 (0.005) | 0.763 (0.099) | 0.155 (0.093) | 0.144 (0.073) | 0.244 (0.044) | 0.096 (0.039) | 0.176 (0.047) |
| FedFB | 0.821 (0.000) | 0.823 (0.001) | 0.823 (0.002) | 0.821 (0.002) | 0.808 (0.007) | 0.064 (0.002) | 0.051 (0.015) | 0.050 (0.009) | 0.045 (0.009) | 0.043 (0.018) |
| BGL | 0.825 (0.001) | 0.826 (0.001) | 0.827 (0.001) | 0.825 (0.001) | 0.818 (0.008) | 0.070 (0.021) | 0.074 (0.006) | 0.084 (0.014) | 0.085 (0.014) | 0.092 (0.022) |
| Tracking | 0.823 (0.008) | 0.825 (0.001) | 0.825 (0.001) | 0.826 (0.002) | 0.818 (0.004) | 0.072 (0.019) | 0.056 (0.002) | 0.056 (0.004) | 0.068 (0.012) | 0.061 (0.005) |
| LoGoFair | 0.699 (0.074) | 0.741 (0.067) | 0.740 (0.084) | 0.643 (0.090) | 0.588 (0.089) | 0.104 (0.103) | 0.040 (0.010) | 0.027 (0.006) | 0.080 (0.050) | 0.066 (0.052) |
| FedFACT-in | 0.564 (0.073) | 0.512 (0.089) | 0.573 (0.051) | – | – | 0.021 (0.008) | 0.018 (0.008) | 0.019 (0.008) | – | – |
| FedFACT-post | 0.826 (0.003) | 0.826 (0.002) | 0.826 (0.002) | – | – | 0.064 (0.011) | 0.052 (0.013) | 0.060 (0.011) | – | – |
| GloReFair | 0.825 (0.000) | 0.824 (0.000) | 0.825 (0.000) | 0.825 (0.002) | 0.826 (0.001) | 0.059 (0.002) | 0.058 (0.004) | 0.054 (0.000) | 0.065 (0.008) | 0.058 (0.004) |

Table 5: The performance of different methods under the EOD criterion across datasets with varying joint-shift severity.

| Method | Adult | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy (↑) | | | | | EOD Gap (↓) | | | | |
| | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ |
| FedAvg | 0.839 (0.000) | 0.830 (0.006) | 0.817 (0.016) | 0.808 (0.028) | 0.804 (0.028) | 0.186 (0.084) | 0.099 (0.041) | 0.211 (0.058) | 0.139 (0.096) | 0.349 (0.234) |
| FairFed | 0.834 (0.003) | 0.810 (0.019) | 0.826 (0.014) | 0.819 (0.021) | 0.790 (0.033) | 0.135 (0.130) | 0.293 (0.176) | 0.184 (0.109) | 0.235 (0.164) | 0.471 (0.124) |
| FedFB | 0.838 (0.001) | 0.829 (0.006) | 0.817 (0.017) | 0.806 (0.026) | 0.806 (0.026) | 0.089 (0.055) | 0.083 (0.019) | 0.145 (0.032) | 0.073 (0.073) | 0.288 (0.140) |
| BGL | 0.797 (0.002) | 0.798 (0.002) | 0.799 (0.000) | 0.802 (0.002) | 0.797 (0.007) | 0.079 (0.006) | 0.101 (0.021) | 0.110 (0.026) | 0.081 (0.007) | 0.139 (0.035) |
| Tracking | 0.833 (0.001) | 0.828 (0.005) | 0.818 (0.012) | 0.808 (0.020) | 0.812 (0.025) | 0.045 (0.002) | 0.036 (0.004) | 0.067 (0.024) | 0.060 (0.036) | 0.102 (0.080) |
| LoGoFair | 0.793 (0.030) | 0.748 (0.018) | 0.729 (0.022) | 0.736 (0.034) | 0.342 (0.169) | 0.112 (0.003) | 0.241 (0.122) | 0.250 (0.178) | 0.287 (0.249) | 0.616 (0.120) |
| FedFACT-in | 0.713 (0.010) | 0.706 (0.006) | 0.637 (0.023) | – | – | 0.050 (0.022) | 0.080 (0.021) | 0.148 (0.059) | – | – |
| FedFACT-post | 0.771 (0.001) | 0.770 (0.014) | 0.771 (0.014) | – | – | 0.154 (0.102) | 0.241 (0.083) | 0.163 (0.070) | – | – |
| GloReFair | 0.832 (0.000) | 0.827 (0.003) | 0.824 (0.004) | 0.819 (0.008) | 0.820 (0.008) | 0.042 (0.005) | 0.069 (0.011) | 0.060 (0.025) | 0.060 (0.032) | 0.092 (0.050) |
| | CreditCard | | | | | | | | | |
| | Accuracy (↑) | | | | | EOD Gap (↓) | | | | |
| | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ |
| FedAvg | 0.796 (0.008) | 0.791 (0.008) | 0.791 (0.005) | 0.796 (0.003) | 0.792 (0.009) | 0.071 (0.034) | 0.136 (0.096) | 0.146 (0.094) | 0.259 (0.044) | 0.144 (0.126) |
| FairFed | 0.796 (0.007) | 0.791 (0.009) | 0.790 (0.004) | 0.795 (0.003) | 0.792 (0.008) | 0.074 (0.044) | 0.129 (0.093) | 0.139 (0.089) | 0.238 (0.044) | 0.138 (0.111) |
| FedFB | 0.795 (0.007) | 0.791 (0.008) | 0.790 (0.004) | 0.794 (0.003) | 0.792 (0.008) | 0.077 (0.045) | 0.125 (0.090) | 0.137 (0.085) | 0.226 (0.047) | 0.134 (0.100) |
| BGL | 0.755 (0.015) | 0.750 (0.020) | 0.742 (0.017) | 0.717 (0.043) | 0.735 (0.073) | 0.022 (0.010) | 0.040 (0.007) | 0.032 (0.013) | 0.091 (0.052) | 0.088 (0.094) |
| Tracking | 0.796 (0.007) | 0.793 (0.010) | 0.791 (0.004) | 0.793 (0.003) | 0.790 (0.007) | 0.053 (0.032) | 0.078 (0.061) | 0.098 (0.059) | 0.165 (0.043) | 0.098 (0.064) |
| LoGoFair | 0.708 (0.068) | 0.667 (0.128) | 0.681 (0.032) | 0.493 (0.127) | 0.727 (0.138) | 0.296 (0.188) | 0.303 (0.267) | 0.241 (0.077) | 0.495 (0.128) | 0.399 (0.309) |
| FedFACT-in | 0.582 (0.040) | 0.603 (0.076) | 0.608 (0.036) | – | – | 0.111 (0.071) | 0.092 (0.085) | 0.068 (0.046) | – | – |
| FedFACT-post | 0.793 (0.006) | 0.791 (0.008) | 0.789 (0.004) | – | – | 0.083 (0.055) | 0.113 (0.087) | 0.126 (0.076) | – | – |
| GloReFair | 0.797 (0.002) | 0.793 (0.006) | 0.793 (0.001) | 0.792 (0.001) | 0.789 (0.005) | 0.039 (0.030) | 0.033 (0.018) | 0.038 (0.027) | 0.047 (0.028) | 0.039 (0.009) |
| | Dutch | | | | | | | | | |
| | Accuracy (↑) | | | | | EOD Gap (↓) | | | | |
| | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ |
| FedAvg | 0.833 (0.000) | 0.830 (0.004) | 0.831 (0.002) | 0.816 (0.019) | 0.806 (0.020) | 0.187 (0.050) | 0.160 (0.094) | 0.180 (0.088) | 0.409 (0.102) | 0.476 (0.087) |
| FairFed | 0.829 (0.004) | 0.817 (0.020) | 0.831 (0.003) | 0.807 (0.025) | 0.797 (0.022) | 0.160 (0.018) | 0.248 (0.150) | 0.247 (0.111) | 0.297 (0.152) | 0.517 (0.173) |
| FedFB | 0.823 (0.003) | 0.814 (0.007) | 0.810 (0.015) | 0.782 (0.050) | 0.733 (0.042) | 0.070 (0.009) | 0.033 (0.013) | 0.038 (0.025) | 0.042 (0.016) | 0.021 (0.018) |
| BGL | 0.825 (0.001) | 0.825 (0.001) | 0.825 (0.000) | 0.826 (0.001) | 0.824 (0.006) | 0.076 (0.013) | 0.075 (0.002) | 0.076 (0.005) | 0.062 (0.003) | 0.078 (0.037) |
| Tracking | 0.829 (0.001) | 0.829 (0.001) | 0.829 (0.003) | 0.830 (0.002) | 0.831 (0.002) | 0.078 (0.003) | 0.084 (0.031) | 0.112 (0.040) | 0.136 (0.040) | 0.167 (0.074) |
| LoGoFair | 0.741 (0.038) | 0.704 (0.093) | 0.701 (0.051) | 0.692 (0.128) | 0.429 (0.218) | 0.149 (0.057) | 0.265 (0.167) | 0.139 (0.121) | 0.147 (0.141) | 0.523 (0.147) |
| FedFACT-in | 0.604 (0.017) | 0.565 (0.020) | – | – | – | 0.101 (0.065) | 0.134 (0.045) | – | – | – |
| FedFACT-post | 0.796 (0.005) | 0.713 (0.088) | – | – | – | 0.118 (0.050) | 0.126 (0.066) | – | – | – |
| GloReFair | 0.825 (0.001) | 0.825 (0.000) | 0.825 (0.001) | 0.825 (0.001) | 0.824 (0.002) | 0.063 (0.005) | 0.058 (0.002) | 0.064 (0.004) | 0.062 (0.004) | 0.060 (0.005) |

Table 6: The performance of different methods under the EO criterion across datasets with varying group-shift severity.

| Method | Adult | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy (↑) | | | | | EO Gap (↓) | | | | |
| | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ |
| FedAvg | 0.840 (0.000) | 0.840 (0.001) | 0.840 (0.001) | 0.841 (0.000) | 0.841 (0.001) | 0.125 (0.004) | 0.122 (0.004) | 0.117 (0.012) | 0.115 (0.011) | 0.122 (0.002) |
| FairFed | 0.838 (0.000) | 0.840 (0.001) | 0.839 (0.001) | 0.839 (0.001) | 0.840 (0.001) | 0.127 (0.007) | 0.127 (0.007) | 0.119 (0.012) | 0.117 (0.020) | 0.122 (0.005) |
| FedFB | 0.824 (0.000) | 0.825 (0.000) | 0.824 (0.001) | 0.826 (0.000) | 0.826 (0.001) | 0.016 (0.010) | 0.024 (0.002) | 0.038 (0.011) | 0.019 (0.009) | 0.036 (0.008) |
| BGL | 0.794 (0.002) | 0.795 (0.000) | 0.797 (0.002) | 0.798 (0.005) | 0.794 (0.004) | 0.007 (0.008) | 0.011 (0.003) | 0.014 (0.005) | 0.021 (0.013) | 0.023 (0.012) |
| Tracking | 0.798 (0.031) | 0.806 (0.024) | 0.782 (0.032) | 0.798 (0.016) | 0.812 (0.019) | 0.058 (0.024) | 0.072 (0.019) | 0.138 (0.130) | 0.073 (0.082) | 0.295 (0.362) |
| LoGoFair | 0.823 (0.004) | 0.813 (0.011) | 0.815 (0.011) | 0.790 (0.020) | 0.705 (0.052) | 0.042 (0.042) | 0.060 (0.022) | 0.061 (0.038) | 0.188 (0.104) | 0.509 (0.236) |
| FedFACT-in | 0.726 (0.027) | 0.737 (0.014) | 0.730 (0.016) | 0.713 (0.024) | – | 0.065 (0.086) | 0.038 (0.039) | 0.057 (0.069) | 0.130 (0.067) | – |
| FedFACT-post | 0.836 (0.000) | 0.836 (0.000) | 0.836 (0.001) | 0.836 (0.000) | – | 0.124 (0.003) | 0.119 (0.006) | 0.118 (0.006) | 0.112 (0.011) | – |
| GloReFair | 0.835 (0.000) | 0.805 (0.021) | 0.785 (0.012) | 0.835 (0.000) | 0.835 (0.000) | 0.056 (0.002) | 0.088 (0.082) | 0.105 (0.069) | 0.054 (0.002) | 0.050 (0.006) |
| | CreditCard | | | | | | | | | |
| | Accuracy (↑) | | | | | EO Gap (↓) | | | | |
| | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ |
| FedAvg | 0.807 (0.001) | 0.805 (0.000) | 0.806 (0.000) | 0.806 (0.000) | 0.807 (0.001) | 0.063 (0.015) | 0.053 (0.014) | 0.056 (0.009) | 0.055 (0.011) | 0.056 (0.014) |
| FairFed | 0.805 (0.001) | 0.805 (0.001) | 0.805 (0.001) | 0.805 (0.000) | 0.805 (0.000) | 0.058 (0.010) | 0.058 (0.017) | 0.060 (0.008) | 0.055 (0.014) | 0.060 (0.018) |
| FedFB | 0.797 (0.002) | 0.795 (0.001) | 0.797 (0.000) | 0.797 (0.001) | 0.799 (0.001) | 0.017 (0.005) | 0.014 (0.001) | 0.018 (0.004) | 0.023 (0.006) | 0.019 (0.003) |
| BGL | 0.560 (0.002) | 0.562 (0.004) | 0.535 (0.025) | 0.550 (0.021) | 0.558 (0.033) | 0.081 (0.014) | 0.074 (0.008) | 0.077 (0.003) | 0.088 (0.012) | 0.096 (0.019) |
| Tracking | 0.804 (0.000) | 0.804 (0.000) | 0.805 (0.001) | 0.805 (0.001) | 0.805 (0.001) | 0.055 (0.010) | 0.047 (0.008) | 0.046 (0.007) | 0.048 (0.008) | 0.051 (0.009) |
| LoGoFair | 0.802 (0.001) | 0.787 (0.016) | 0.655 (0.181) | 0.703 (0.117) | 0.722 (0.062) | 0.029 (0.010) | 0.055 (0.040) | 0.100 (0.087) | 0.232 (0.186) | 0.101 (0.066) |
| FedFACT-in | 0.744 (0.013) | 0.743 (0.010) | 0.729 (0.026) | – | – | 0.017 (0.002) | 0.016 (0.003) | 0.031 (0.026) | – | – |
| FedFACT-post | 0.802 (0.001) | 0.802 (0.000) | 0.803 (0.001) | – | – | 0.061 (0.012) | 0.051 (0.007) | 0.056 (0.005) | – | – |
| GloReFair | 0.803 (0.001) | 0.803 (0.000) | 0.803 (0.000) | 0.803 (0.001) | 0.803 (0.000) | 0.041 (0.005) | 0.038 (0.002) | 0.043 (0.005) | 0.042 (0.004) | 0.042 (0.006) |
| | Dutch | | | | | | | | | |
| | Accuracy (↑) | | | | | EO Gap (↓) | | | | |
| | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ |
| FedAvg | 0.834 (0.001) | 0.835 (0.001) | 0.834 (0.002) | 0.835 (0.002) | 0.835 (0.000) | 0.090 (0.004) | 0.085 (0.008) | 0.082 (0.013) | 0.084 (0.012) | 0.088 (0.004) |
| FairFed | 0.834 (0.001) | 0.836 (0.001) | 0.835 (0.001) | 0.835 (0.001) | 0.836 (0.000) | 0.088 (0.002) | 0.088 (0.006) | 0.089 (0.008) | 0.087 (0.008) | 0.091 (0.002) |
| FedFB | 0.833 (0.001) | 0.833 (0.001) | 0.833 (0.001) | 0.834 (0.001) | 0.833 (0.001) | 0.057 (0.004) | 0.053 (0.005) | 0.050 (0.011) | 0.049 (0.007) | 0.051 (0.003) |
| BGL | 0.760 (0.008) | 0.751 (0.002) | 0.750 (0.003) | 0.763 (0.017) | 0.768 (0.016) | 0.082 (0.006) | 0.076 (0.001) | 0.073 (0.003) | 0.080 (0.014) | 0.083 (0.014) |
| Tracking | 0.522 (0.036) | 0.546 (0.026) | 0.614 (0.108) | 0.611 (0.060) | 0.632 (0.071) | 0.014 (0.020) | 0.005 (0.007) | 0.045 (0.063) | 0.460 (0.412) | 0.363 (0.258) |
| LoGoFair | 0.830 (0.001) | 0.828 (0.004) | 0.789 (0.034) | 0.811 (0.016) | 0.676 (0.148) | 0.011 (0.002) | 0.039 (0.035) | 0.024 (0.011) | 0.065 (0.069) | 0.337 (0.113) |
| FedFACT-in | 0.593 (0.016) | 0.578 (0.019) | – | – | – | 0.027 (0.019) | 0.031 (0.016) | – | – | – |
| FedFACT-post | 0.832 (0.000) | 0.832 (0.000) | – | – | – | 0.002 (0.003) | 0.003 (0.003) | – | – | – |
| GloReFair | 0.830 (0.001) | 0.830 (0.001) | 0.830 (0.000) | 0.831 (0.001) | 0.830 (0.000) | 0.015 (0.006) | 0.017 (0.005) | 0.017 (0.001) | 0.020 (0.012) | 0.013 (0.009) |

Table 7: The performance of different methods under the EO criterion across datasets with varying label-shift severity.

| Method | Adult | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy (↑) | | | | | EO Gap (↓) | | | | |
| | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ |
| FedAvg | 0.834 (0.007) | 0.833 (0.003) | 0.810 (0.035) | 0.803 (0.036) | 0.790 (0.020) | 0.090 (0.034) | 0.085 (0.009) | 0.052 (0.033) | 0.040 (0.029) | 0.030 (0.009) |
| FairFed | 0.833 (0.006) | 0.833 (0.004) | 0.816 (0.032) | 0.803 (0.036) | 0.772 (0.043) | 0.124 (0.023) | 0.099 (0.040) | 0.100 (0.068) | 0.051 (0.042) | 0.070 (0.057) |
| FedFB | 0.818 (0.004) | 0.817 (0.005) | 0.801 (0.034) | 0.796 (0.032) | 0.807 (0.008) | 0.051 (0.021) | 0.050 (0.017) | 0.026 (0.008) | 0.028 (0.029) | 0.083 (0.045) |
| BGL | 0.780 (0.023) | 0.779 (0.028) | 0.785 (0.032) | 0.757 (0.052) | 0.799 (0.035) | 0.041 (0.021) | 0.034 (0.015) | 0.039 (0.026) | 0.069 (0.030) | 0.081 (0.036) |
| Tracking | 0.815 (0.001) | 0.818 (0.005) | 0.796 (0.031) | 0.797 (0.032) | 0.811 (0.005) | 0.016 (0.012) | 0.033 (0.022) | 0.021 (0.012) | 0.031 (0.041) | 0.027 (0.007) |
| LoGoFair | 0.741 (0.042) | 0.731 (0.030) | 0.683 (0.077) | 0.498 (0.106) | 0.668 (0.197) | 0.039 (0.018) | 0.012 (0.007) | 0.028 (0.014) | 0.028 (0.010) | 0.099 (0.040) |
| FedFACT-in | 0.663 (0.031) | 0.680 (0.080) | 0.594 (0.089) | 0.457 (0.104) | – | 0.009 (0.007) | 0.010 (0.006) | 0.009 (0.007) | 0.019 (0.003) | – |
| FedFACT-post | 0.831 (0.005) | 0.830 (0.002) | 0.807 (0.035) | 0.803 (0.036) | – | 0.098 (0.017) | 0.097 (0.006) | 0.063 (0.036) | 0.057 (0.039) | – |
| GloReFair | 0.832 (0.004) | 0.831 (0.003) | 0.822 (0.018) | 0.817 (0.018) | 0.814 (0.009) | 0.040 (0.018) | 0.057 (0.012) | 0.062 (0.011) | 0.033 (0.013) | 0.022 (0.017) |
| | CreditCard | | | | | | | | | |
| | Accuracy (↑) | | | | | EO Gap (↓) | | | | |
| | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ |
| FedAvg | 0.800 (0.004) | 0.795 (0.011) | 0.793 (0.010) | 0.788 (0.007) | 0.788 (0.007) | 0.049 (0.007) | 0.036 (0.023) | 0.041 (0.030) | 0.046 (0.032) | 0.044 (0.030) |
| FairFed | 0.803 (0.006) | 0.799 (0.011) | 0.795 (0.009) | 0.793 (0.013) | 0.793 (0.010) | 0.044 (0.012) | 0.048 (0.013) | 0.051 (0.020) | 0.041 (0.031) | 0.042 (0.036) |
| FedFB | 0.787 (0.002) | 0.787 (0.006) | 0.788 (0.006) | 0.785 (0.005) | 0.788 (0.005) | 0.010 (0.001) | 0.003 (0.004) | 0.004 (0.003) | 0.006 (0.003) | 0.011 (0.005) |
| BGL | 0.550 (0.055) | 0.609 (0.127) | 0.613 (0.133) | 0.650 (0.138) | 0.471 (0.069) | 0.073 (0.003) | 0.067 (0.007) | 0.068 (0.002) | 0.075 (0.013) | 0.063 (0.011) |
| Tracking | 0.799 (0.002) | 0.794 (0.010) | 0.793 (0.010) | 0.790 (0.008) | 0.791 (0.009) | 0.038 (0.006) | 0.035 (0.019) | 0.035 (0.019) | 0.036 (0.024) | 0.044 (0.030) |
| LoGoFair | 0.732 (0.067) | 0.786 (0.020) | 0.787 (0.039) | 0.726 (0.116) | 0.383 (0.112) | 0.009 (0.004) | 0.032 (0.017) | 0.023 (0.003) | 0.048 (0.042) | 0.013 (0.003) |
| FedFACT-in | 0.660 (0.061) | 0.628 (0.072) | 0.687 (0.062) | – | – | 0.013 (0.005) | 0.018 (0.009) | 0.008 (0.003) | – | – |
| FedFACT-post | 0.795 (0.002) | 0.792 (0.008) | 0.792 (0.009) | – | – | 0.051 (0.005) | 0.039 (0.023) | 0.040 (0.022) | – | – |
| GloReFair | 0.798 (0.001) | 0.795 (0.008) | 0.794 (0.008) | 0.791 (0.006) | 0.791 (0.006) | 0.043 (0.003) | 0.041 (0.006) | 0.039 (0.005) | 0.038 (0.008) | 0.038 (0.007) |
| | Dutch | | | | | | | | | |
| | Accuracy (↑) | | | | | EO Gap (↓) | | | | |
| | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ |
| FedAvg | 0.831 (0.003) | 0.834 (0.001) | 0.836 (0.001) | 0.832 (0.006) | 0.831 (0.007) | 0.072 (0.016) | 0.084 (0.014) | 0.095 (0.010) | 0.077 (0.040) | 0.069 (0.043) |
| FairFed | 0.830 (0.006) | 0.836 (0.001) | 0.834 (0.003) | 0.833 (0.001) | 0.748 (0.063) | 0.065 (0.039) | 0.103 (0.004) | 0.110 (0.008) | 0.087 (0.020) | 0.106 (0.041) |
| FedFB | 0.831 (0.002) | 0.833 (0.002) | 0.835 (0.002) | 0.832 (0.006) | 0.829 (0.006) | 0.040 (0.013) | 0.053 (0.018) | 0.065 (0.016) | 0.051 (0.036) | 0.063 (0.044) |
| BGL | 0.754 (0.051) | 0.742 (0.031) | 0.736 (0.029) | 0.755 (0.052) | 0.692 (0.031) | 0.081 (0.018) | 0.076 (0.014) | 0.073 (0.008) | 0.076 (0.019) | 0.054 (0.009) |
| Tracking | 0.670 (0.078) | 0.655 (0.094) | 0.590 (0.081) | 0.657 (0.098) | 0.669 (0.057) | 0.294 (0.208) | 0.456 (0.412) | 0.287 (0.403) | 0.366 (0.426) | 0.508 (0.077) |
| LoGoFair | 0.738 (0.052) | 0.759 (0.058) | 0.759 (0.081) | 0.673 (0.088) | 0.642 (0.065) | 0.091 (0.114) | 0.027 (0.006) | 0.019 (0.007) | 0.007 (0.003) | 0.013 (0.008) |
| FedFACT-in | 0.500 (0.091) | 0.491 (0.098) | 0.554 (0.068) | – | 0.449 (0.100) | 0.008 (0.006) | 0.017 (0.007) | 0.010 (0.010) | – | 0.008 (0.008) |
| FedFACT-post | 0.828 (0.004) | 0.830 (0.003) | 0.829 (0.002) | – | 0.830 (0.005) | 0.028 (0.006) | 0.014 (0.004) | 0.016 (0.008) | – | 0.020 (0.005) |
| GloReFair | 0.830 (0.001) | 0.830 (0.000) | 0.830 (0.001) | 0.830 (0.002) | 0.829 (0.001) | 0.015 (0.005) | 0.015 (0.006) | 0.018 (0.001) | 0.017 (0.015) | 0.013 (0.007) |

Table 8: The performance of different methods under the EO criterion across datasets with varying joint-shift severity.

| Method | Adult | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy (↑) | | | | | EO Gap (↓) | | | | |
| | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ |
| FedAvg | 0.839 (0.000) | 0.830 (0.006) | 0.817 (0.016) | 0.808 (0.028) | 0.804 (0.028) | 0.186 (0.084) | 0.099 (0.041) | 0.211 (0.058) | 0.139 (0.096) | 0.349 (0.234) |
| FairFed | 0.832 (0.005) | 0.810 (0.019) | 0.823 (0.015) | 0.834 (0.004) | 0.792 (0.032) | 0.133 (0.149) | 0.282 (0.198) | 0.210 (0.169) | 0.244 (0.169) | 0.493 (0.150) |
| FedFB | 0.821 (0.003) | 0.818 (0.004) | 0.815 (0.004) | 0.795 (0.018) | 0.802 (0.017) | 0.036 (0.021) | 0.043 (0.030) | 0.059 (0.026) | 0.117 (0.076) | 0.096 (0.059) |
| BGL | 0.775 (0.022) | 0.793 (0.018) | 0.797 (0.020) | 0.787 (0.038) | 0.777 (0.041) | 0.037 (0.037) | 0.081 (0.065) | 0.110 (0.066) | 0.081 (0.047) | 0.154 (0.080) |
| Tracking | 0.821 (0.004) | 0.816 (0.009) | 0.813 (0.009) | 0.776 (0.019) | 0.776 (0.030) | 0.044 (0.031) | 0.038 (0.020) | 0.044 (0.024) | 0.049 (0.054) | 0.013 (0.013) |
| LoGoFair | 0.734 (0.089) | 0.751 (0.020) | 0.721 (0.006) | 0.685 (0.031) | 0.586 (0.140) | 0.157 (0.121) | 0.106 (0.111) | 0.205 (0.218) | 0.358 (0.255) | 0.186 (0.205) |
| FedFACT-in | 0.743 (0.009) | 0.716 (0.014) | 0.683 (0.040) | – | – | 0.034 (0.022) | 0.110 (0.054) | 0.205 (0.185) | – | – |
| FedFACT-post | 0.836 (0.001) | 0.828 (0.005) | 0.818 (0.011) | – | – | 0.168 (0.064) | 0.085 (0.056) | 0.152 (0.100) | – | – |
| GloReFair | 0.834 (0.001) | 0.830 (0.007) | 0.826 (0.009) | 0.821 (0.005) | 0.821 (0.009) | 0.060 (0.014) | 0.061 (0.027) | 0.083 (0.031) | 0.033 (0.026) | 0.088 (0.051) |
| | CreditCard | | | | | | | | | |
| | Accuracy (↑) | | | | | EO Gap (↓) | | | | |
| | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ |
| FedAvg | 0.796 (0.008) | 0.791 (0.008) | 0.791 (0.005) | 0.796 (0.003) | 0.792 (0.009) | 0.071 (0.034) | 0.136 (0.096) | 0.146 (0.094) | 0.259 (0.044) | 0.144 (0.126) |
| FairFed | 0.803 (0.005) | 0.791 (0.008) | 0.792 (0.006) | 0.793 (0.001) | 0.794 (0.013) | 0.165 (0.067) | 0.195 (0.132) | 0.157 (0.096) | 0.241 (0.085) | 0.055 (0.063) |
| FedFB | 0.787 (0.004) | 0.785 (0.005) | 0.782 (0.001) | 0.781 (0.001) | 0.782 (0.003) | 0.007 (0.002) | 0.007 (0.005) | 0.012 (0.001) | 0.010 (0.003) | 0.003 (0.002) |
| BGL | 0.587 (0.060) | 0.572 (0.038) | 0.545 (0.019) | 0.492 (0.066) | 0.599 (0.181) | 0.083 (0.039) | 0.084 (0.036) | 0.105 (0.032) | 0.134 (0.027) | 0.105 (0.086) |
| Tracking | 0.796 (0.007) | 0.792 (0.009) | 0.790 (0.004) | 0.793 (0.002) | 0.791 (0.007) | 0.066 (0.045) | 0.101 (0.078) | 0.120 (0.074) | 0.191 (0.048) | 0.116 (0.071) |
| LoGoFair | 0.703 (0.071) | 0.690 (0.137) | 0.599 (0.111) | 0.519 (0.210) | 0.516 (0.195) | 0.154 (0.096) | 0.205 (0.207) | 0.145 (0.164) | 0.194 (0.141) | 0.151 (0.154) |
| FedFACT-in | 0.703 (0.033) | 0.646 (0.094) | 0.638 (0.059) | – | – | 0.076 (0.083) | 0.127 (0.176) | 0.066 (0.025) | – | – |
| FedFACT-post | 0.793 (0.006) | 0.791 (0.008) | 0.789 (0.004) | – | – | 0.083 (0.055) | 0.113 (0.087) | 0.126 (0.076) | – | – |
| GloReFair | 0.797 (0.003) | 0.794 (0.006) | 0.793 (0.001) | 0.792 (0.002) | 0.791 (0.005) | 0.044 (0.039) | 0.048 (0.027) | 0.056 (0.039) | 0.091 (0.015) | 0.067 (0.029) |
| | Dutch | | | | | | | | | |
| | Accuracy (↑) | | | | | EO Gap (↓) | | | | |
| | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ | $H = 1.0$ | $H = 0.8$ | $H = 0.6$ | $H = 0.4$ | $H = 0.2$ |
| FedAvg | 0.833 (0.000) | 0.830 (0.004) | 0.831 (0.002) | 0.816 (0.019) | 0.806 (0.020) | 0.087 (0.045) | 0.092 (0.048) | 0.084 (0.066) | 0.317 (0.168) | 0.356 (0.166) |
| FairFed | 0.832 (0.002) | 0.822 (0.018) | 0.829 (0.006) | 0.811 (0.032) | 0.776 (0.016) | 0.093 (0.040) | 0.242 (0.159) | 0.120 (0.050) | 0.255 (0.143) | 0.374 (0.161) |
| FedFB | 0.832 (0.002) | 0.830 (0.004) | 0.830 (0.003) | 0.817 (0.020) | 0.806 (0.022) | 0.055 (0.046) | 0.080 (0.044) | 0.062 (0.067) | 0.302 (0.169) | 0.345 (0.160) |
| BGL | 0.763 (0.022) | 0.796 (0.001) | 0.794 (0.012) | 0.777 (0.030) | 0.785 (0.054) | 0.076 (0.013) | 0.100 (0.008) | 0.092 (0.016) | 0.119 (0.022) | 0.096 (0.013) |
| Tracking | 0.666 (0.044) | 0.610 (0.116) | 0.724 (0.031) | 0.679 (0.108) | 0.576 (0.069) | 0.919 (0.079) | 0.031 (0.044) | 0.394 (0.124) | 0.247 (0.260) | 0.173 (0.245) |
| LoGoFair | 0.781 (0.053) | 0.754 (0.039) | 0.724 (0.036) | 0.661 (0.125) | 0.595 (0.161) | 0.073 (0.032) | 0.110 (0.116) | 0.164 (0.149) | 0.074 (0.068) | 0.190 (0.130) |
| FedFACT-in | 0.533 (0.031) | 0.547 (0.014) | – | – | – | 0.160 (0.179) | 0.167 (0.062) | – | – | – |
| FedFACT-post | 0.827 (0.003) | 0.795 (0.022) | – | – | – | 0.060 (0.041) | 0.097 (0.058) | – | – | – |
| GloReFair | 0.830 (0.001) | 0.829 (0.001) | 0.829 (0.001) | 0.829 (0.001) | 0.827 (0.002) | 0.010 (0.008) | 0.008 (0.008) | 0.019 (0.016) | 0.039 (0.020) | 0.048 (0.017) |

### H.3 Explanation and analysis of ablation experiment.

The EOD and EO ablation tables in the main paper examine three components of GloReFair under joint Dirichlet heterogeneity ($H = 0.2$). Within each ablation experiment, the variants share the same data splits, model, hyperparameters, and random seeds.

**Ablation variants.**

- **Local-ratio.** This variant replaces pooled channel rates with equal-weight averages of client-local rates. Let $n_{k,g,y}$ be the number of samples in client $k$'s channel $(g, y)$, and let $R_{k,g,y}$ denote either the mean binary prediction or the mean predicted probability in that channel, evaluated at $\theta^{t+1}$. For each globally active channel, the server uses

$$R^{\mathrm{loc}}_{g,y} = \frac{1}{|\mathcal{K}_{g,y}|} \sum_{k \in \mathcal{K}_{g,y}} R_{k,g,y}, \qquad \mathcal{K}_{g,y} = \{k : n_{k,g,y} > 0\}.$$

  Locally empty channels are excluded. Group-reference means, residual construction, and subsequent updates follow GloReFair.
- **Coarse.** This variant ties the two label-specific prices within each group, $\lambda_{g,0} = \lambda_{g,1}$, replacing separate group-label control with group-level control. It tests whether EOD benefits from independently adjusting positive- and negative-label sample weights. Coarse is omitted under EO, where only the positive-label channel is updated and negative-label prices remain zero.
- **Frozen-$\Lambda$.** This variant disables adaptive price updates by keeping $\Lambda^{t+1} = \Lambda^t$ during training. Sample weights are still recomputed from the losses at each synchronized model, so the ablation isolates adaptation of channel prices to fairness feedback while retaining loss-dependent reweighting.

**Results and analysis.**

- **Pooled statistics improve fairness feedback.** Compared with Local-ratio, GloReFair reduces the Dutch EOD gap from 0.094 to 0.060 and the Adult EO gap from 0.187 to 0.088, with respective accuracies of 0.824 versus 0.822 and 0.821 versus 0.816. Pooling weights each local channel rate by its sample count, whereas Local-ratio gives clients with small and large channel support equal influence. The pooled feedback therefore matches the fairness rates of the pooled data.
- **Separate label-channel prices improve EOD control.** Coarse increases the EOD gap from 0.092 to 0.172 on Adult, from 0.039 to 0.115 on CreditCard, and from 0.060 to 0.342 on Dutch. Tying prices restricts the controller's ability to respond separately to TPR deficits and FPR excesses within a group. The results are consistent with the expressivity distinction in Appendix E.
- **Adaptive prices improve both EO and EOD control.** Frozen-$\Lambda$ produces larger mean fairness gaps than GloReFair on all three datasets under both criteria, while accuracy differs by at most 0.007. Thus, retaining loss-dependent weights with fixed channel prices does not recover the fairness performance of adaptive residual-driven updates in these settings.

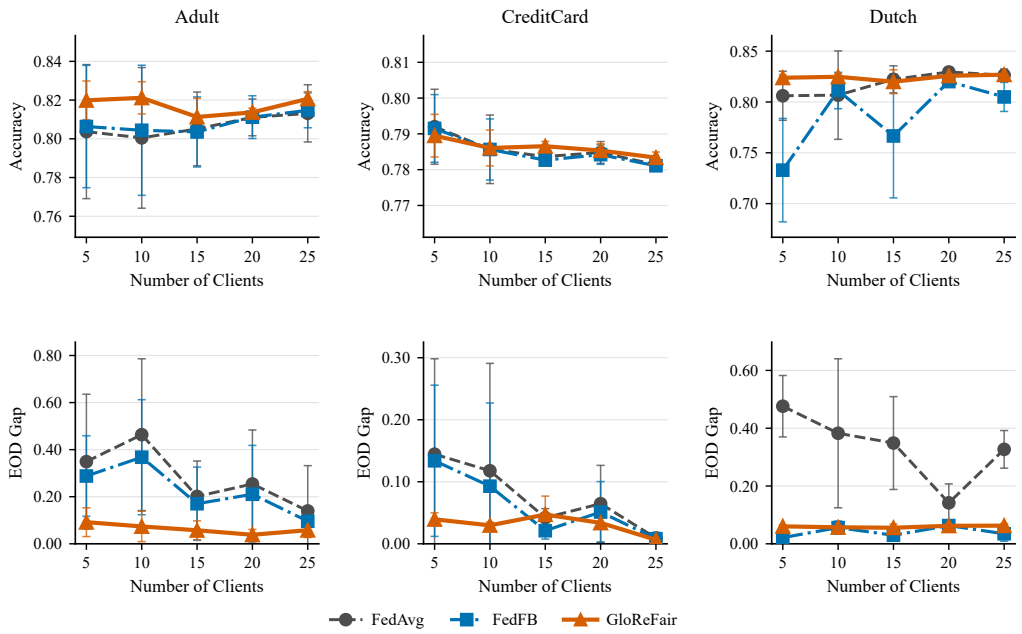


Figure 5: Client-count sensitivity under EOD. The top row shows accuracy and the bottom row shows EOD gap. Points are means over three seeds; error bars show one sample standard deviation.

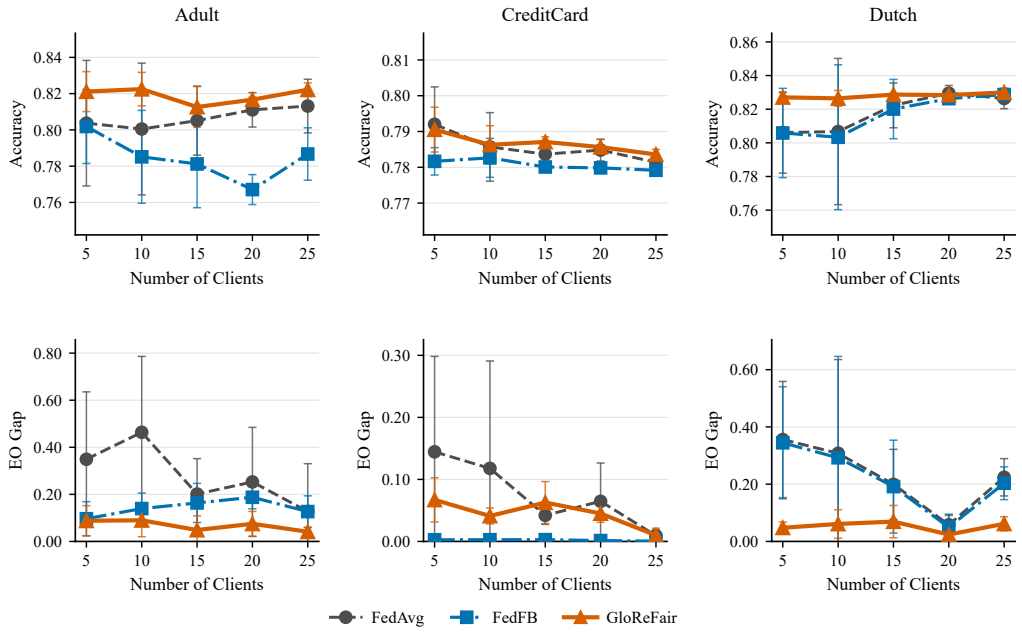


Figure 6: Client-count sensitivity under EO. The top row shows accuracy and the bottom row shows EO gap. Points are means over three seeds; error bars show one sample standard deviation.

### H.4 CLIENT-COUNT SENSITIVITY

Figures 5 and 6 report the performance of GloReFair, FedAvg and FedFB across in scenarios with different client count and under joint Dirichlet heterogeneity ($H = 0.2$). We select FedAvg as the classic FL method, and select FedFB because it perform more excellently comparing to the other fairness-aware baselines in most heterogeneous settings.

With a fixed total training set, increasing the number of clients reduces the average local sample size and can leave individual group-label combinations sparsely represented or absent. This can exacerbate empirical heterogeneity across clients and make local fairness estimates less reliable, posing additional challenges to fairness-aware federated learning. Despite these challenges, GloReFair maintains relatively stable accuracy and low fairness gaps as the number of clients increases from 5 to 25. Its advantage over FedFB is particularly evident on Adult, where it achieves lower fairness gaps under both EOD and EO while maintaining higher accuracy. On Dutch, GloReFair achieves consistently higher and more stable accuracy under EOD, although FedFB often attains smaller fairness gaps; under EO, GloReFair achieves lower gaps with comparable or higher accuracy. These results highlight GloReFair's ability to preserve a favorable balance between predictive performance and fairness as local data become more fragmented. This robustness is consistent with its use of globally aggregated group-label statistics and shared channel prices: fairness feedback is computed from pooled statistics, allowing local reweighting to be coordinated even when individual clients have sparse or incomplete group-label coverage.