A LONG-TAIL PERSPECTIVE ON GROUP FAIRNESS: FAIR LOGIT ADJUSTMENT VIA PRIOR CORRECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning classifiers often exhibit unequal predictive performance across demographic groups. Post-processing offers a practical way to reduce these disparities without retraining, typically by adjusting decision thresholds or prediction distributions. To connect these interventions to the origins of disparity, we examine how the data distribution shapes predictions across groups. A class that is common overall can still be rare within a particular group, exposing a long tail at the group–class level. Under standard risk minimization, errors on these rare combinations receive less weight, while group-conditional priors absorbed into decision scores favor prevalent classes, contributing to unequal recognition across groups. These observations motivate giving each group's recognition performance on each class equal importance. We express this objective through Group-Balanced Error Rate (GBER) and derive its Bayes-optimal rule, which removes the group-conditional prior from posterior scores. This leads to **Fair Logit Adjustment (FLA)**, a deterministic, metric-agnostic post-processing method that directly corrects learned logits. This simple score adjustment requires no constrained solver and naturally supports multi-class and multi-group settings. Across multiple benchmarks and model families, FLA achieves highly competitive fairness–utility trade-offs compared with state-of-the-art post-processing methods, with substantially lower construction cost. The same prior-correction principle also extends to an in-processing formulation. By identifying a distributional source of disparity and the score component to correct, FLA offers an interpretable and computationally lightweight approach to post-processing fairness.

1 INTRODUCTION

Machine learning models deployed in high-stakes domains—hiring, lending, clinical triage, criminal-justice risk assessment—have been repeatedly shown to produce decisions that differ substantially across demographic groups (Barocas & Selbst, 2016; Bolukbasi et al., 2016; Berk et al., 2021). Mitigating these disparities is challenging because fairness interventions typically come with substantial cost: in-processing methods modify the training loop, often the model architecture, and require expensive retraining; post-processing methods—the natural decoupled alternative—avoid retraining, yet can introduce substantial computational and operational costs.

Many existing post-processing methods share a common paradigm: fairness is imposed as an external constraint on the classifier's output space—aligning marginal distributions via optimal transport (Xian et al., 2023) or randomizing group-specific decision thresholds (Hardt et al., 2016). These approaches yield sharp guarantees for specific criteria, with the intervention determined by the desired fairness properties of the outputs.

Departing from this output-centric perspective, we examine how the distribution of *group–class combinations* influences learned decision scores and contributes to group-conditional disparity. A class that is common overall can be rare within a particular protected group, so class marginals can conceal imbalance within groups (Sagawa et al., 2020). Occupation classification provides a concrete example. On BiasBios (De-Arteaga et al., 2019), the ERM classifier in Figure 1(a) has higher recall for women in occupations such as nurse and dietitian, but higher recall for men in occupations such as surgeon and pastor. These recall differences align with the underlying group-conditional class frequencies: each of these occupations is recognized more reliably in the group where it has the

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.

higher class prior. The direction of the gap changes with the occupation: the disadvantaged units are particular group–class cells. This structure motivates examining how the frequency of each class within a group enters the classifier's decisions.

Long-tail learning provides a basis for this analysis. Prior work shows that imbalanced learning favors frequent classes (Xu et al., 2021) and that adjusting a classifier can improve tail recognition with its representation held fixed (Kang et al., 2020; Menon et al., 2020). In our setting, ordinary risk minimization weights each group–class error by its frequency. At the score level, posterior probabilities reflect both the evidence for a class and its prior within the group. Learned logits that approximate these posterior scores therefore give prevalent classes a larger prior contribution. This *prior absorption* identifies a component of decision scores through which group–class imbalance can contribute to unequal recognition.

For group fairness, the importance of correctly recognizing a class should not depend on how frequently it occurs within each group. We therefore give each group's performance on each true class equal importance. Averaging these conditional errors gives *Group-Balanced Error Rate* (GBER), which retains correct classification as its goal while giving rare cells the same weight as prevalent ones. Its Bayes rule compares class-conditional evidence within the observed group; in posterior-score form, this amounts to subtracting the group-conditional log-prior. Classical logit adjustment (Menon et al., 2020) seeks balanced recognition across classes, whereas GBER gives equal importance to each group's recognition performance within each class. This distinction matters because good class-level performance can conceal poor recognition within particular groups. The resulting Bayes rules share the same prior-correction principle, but the GBER objective leads to group-conditional rather than marginal class priors. We also relate GBER to group recall disparities, providing further support for this fairness-motivated objective.

From the Bayes rule of GBER, we derive *Fair Logit Adjustment* (FLA), a closed-form post-processing method that corrects learned logits for group-conditional class priors. The intervention follows from the distributional structure underlying group–class recognition, giving FLA a metric-agnostic correction form without optimizing a prescribed fairness disparity. Returning to the BiasBios example in Figure 1, this correction addresses the opposite patterns of under-recognition within the same classifier: recall improves for men in occupations such as nurse and dietitian, and for women in occupations such as surgeon and pastor, narrowing the gaps in both directions. The distributional perspective thus identifies both the source of the score adjustment and how it can benefit the groups disadvantaged in different classes. FLA implements this intervention deterministically, without a constrained solver or auxiliary model, and naturally supports multiple classes and groups. The same prior-correction principle also extends to a training-loss instantiation when retraining is available.

Contributions.

1. **A long-tail perspective on group unfairness.** We view group unfairness as a long-tail problem over group–class combinations. By connecting group-conditional imbalance to prior absorption in decision scores, this perspective explains how class-level performance can conceal unequal recognition across groups and identifies a distributional basis for fairness intervention.

2. **Group-Balanced Error Rate and Fair Logit Adjustment.** We introduce Group-Balanced Error Rate (GBER) to give each group's recognition performance on each class equal importance, regardless of its prevalence. From its Bayes-optimal rule, we derive Fair Logit Adjustment (FLA), an interpretable and deterministic post-processing method that corrects group-conditional priors in learned logits.

3. **Comprehensive empirical validation of FLA.** We evaluate FLA across multiple benchmarks and model families, including multi-class and multi-group settings. The results demonstrate highly competitive fairness–utility trade-offs against state-of-the-art post-processing methods, with orders-of-magnitude lower construction cost than LP-based approaches and negligible prediction overhead.

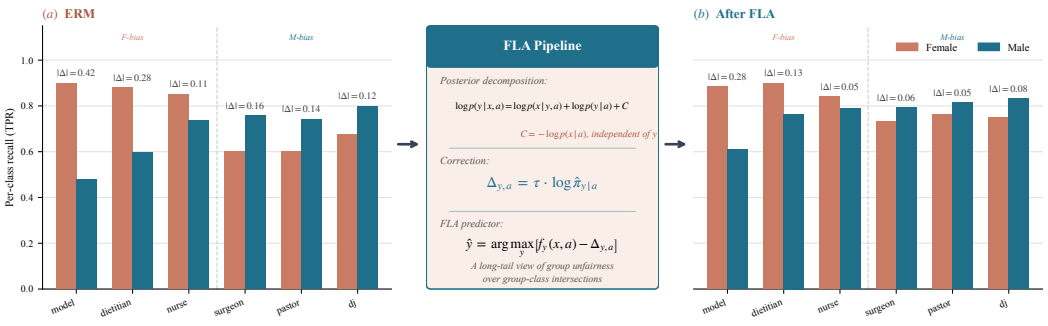

Figure 1: **FLA mechanism and correction (BiasBios BERT).** Six occupations spanning both bias directions: *F-bias* classes (model, dietitian, nurse) where ERM over-recalls Female; *M-bias* classes (surgeon, pastor, dj) where ERM over-recalls Male. *(a) ERM*: large bidirectional gender TPR gaps induced by group-conditional prior terms absorbed in the logits. *Center*: the FLA correction applies the analytic offset $-\tau \log \hat{\pi}_{y|a}$ to each logit, removing the absorbed group-conditional prior without retraining. *(b) After* FLA: gaps reduce in both directions.

2 Related Work

Fairness objectives and post-processing. Fairness can be incorporated into training through adversarial learning (Zhang et al., 2018) or cost-sensitive reductions (Agarwal et al., 2018). Post-processing instead adapts an existing predictor through group-specific thresholds (Hardt et al., 2016), optimal transport (Xian et al., 2023), information projection (Alghamdi et al., 2022), or learned score modification (Tifrea et al., 2023; Chen et al., 2024; Di Gennaro et al., 2024). These approaches offer different ways to implement a desired fairness–accuracy trade-off. Their theoretical foundations include cost-sensitive characterizations of fair classification (Menon & Williamson, 2018) and Bayes-optimal rules under explicit disparity constraints (Zeng et al., 2026). Unconstrained risk minimization also has its own fairness implications: Liu et al. (2019) study how group calibration can arise from learning while other criteria remain violated. Our starting point is the weighting of conditional recognition errors. GBER gives these errors equal importance across group–class cells, and its Bayes rule identifies a prior correction implementable directly on learned logits.

Long-tail learning and prior correction. The preference for frequent classes motivates training rebalancing (Ren et al., 2020) and classifier adjustment. Kang et al. (2020) show that modifying the classifier while keeping representations fixed can improve long-tail recognition. Attribute-aware adjustment (Zhou et al., 2025) and multiplicative corrections (Hasegawa & Sato, 2024) further study imbalance across categories and decision scores. Menon et al. (2020) give a statistical account of additive logit adjustment through class-balanced error. Inspired by their logit-adjustment principle, FLA brings prior correction to group fairness while retaining the same elegant additive form. The recognition objective now gives equal importance to each group's performance on each class, including disparities hidden by class marginals. GBER formalizes this objective, and its Bayes rule leads to group-conditional rather than marginal class priors in the adjustment. This connects the simplicity of logit adjustment to the group–class structure underlying unequal recognition.

3 Fair Logit Adjustment

We formalize the group–class long-tail perspective through an equal-weight recognition objective and derive the prior correction underlying our method.

3.1 Problem Setup

Notation. Let $X \in \mathcal{X}$ denote the input feature, $A \in \mathcal{A}$ the sensitive attribute, and $Y \in \mathcal{Y} = \{1, \ldots, K\}$ the class label ($K \geq 2$, $|\mathcal{A}| = G$). We write lowercase (x, a, y) for their realizations. We adopt the standard fairness post-processing setting in which the sensitive attribute is observed at decision time; the classifier therefore takes both X and A as input, with class-y logit denoted by

$f_y(x, a)$. We use $\pi_{y,a} = \mathbb{P}(Y = y, A = a)$ and $\pi_{y|a} = \mathbb{P}(Y = y \mid A = a)$ for label-prior quantities, $\mathbb{P}$ for probabilities, and p for probability mass functions or densities, as appropriate. In particular, $p(y \mid x, a)$ and $p(y \mid a)$ denote the conditional probabilities of class y. Throughout we assume positivity ($\pi_{y,a} > 0$), joint measurability of (X, Y, A), and the existence of a measurable selection of the pointwise GBER minimizer; full statements are given in Appendix A.

Fairness disparities. For a classifier h, let $\mathrm{TPR}_{y,a} = \mathbb{P}(h(X, A) = y \mid Y = y, A = a)$ denote the per-group recall on class y. We use three standard group-fairness disparities:

$$\Delta\mathrm{EOpp}(h, y) = \max_{a,a'} \left|\mathrm{TPR}_{y,a} - \mathrm{TPR}_{y,a'}\right|,$$

$$\Delta\mathrm{EO}(h) = \max_{y,k} \max_{a,a'} \left|\mathbb{P}(h(X, A) = k \mid Y = y, A = a) - \mathbb{P}(h(X, A) = k \mid Y = y, A = a')\right|,$$

$$\Delta\mathrm{SP}(h) = \max_{y} \max_{a,a'} \left|\mathbb{P}(h(X, A) = y \mid A = a) - \mathbb{P}(h(X, A) = y \mid A = a')\right|.$$

$\Delta\mathrm{EOpp}$ measures per-class recall gap (equalized opportunity (Hardt et al., 2016)); $\Delta\mathrm{EO}$ compares conditional prediction rates across groups for all true and predicted classes (equalized odds); $\Delta\mathrm{SP}$ measures the largest inter-group gap in class prediction rates (statistical parity). For the experimental EOpp summary, we report the positive-class gap in binary tasks and $\max_y \Delta\mathrm{EOpp}(h, y)$ in multi-class tasks.

3.2 A LONG-TAIL PERSPECTIVE ON THE FAIRNESS OBJECTIVE

A rare group–class cell contributes little to aggregate risk even when its members are poorly recognized. For group fairness, its recognition performance should not matter less simply because it is less prevalent. The long-tail perspective therefore motivates assigning equal importance to the conditional recognition errors of all group–class cells, while retaining correct classification as the goal.

From frequency-weighted risk to GBER. Let $e_{y,a}(h) = \mathbb{P}(h(X, A) \neq y \mid Y = y, A = a)$. The population classification risk underlying ordinary error minimization is

$$\mathcal{R}(h) = \mathbb{P}(h(X, A) \neq Y) = \sum_{y,a} \pi_{y,a}\, e_{y,a}(h).$$

The same decomposition applies to a surrogate loss using its conditional expectation in each cell. Thus an equal reduction in conditional error contributes less to the objective when it occurs in a rarer cell. Replacing these frequency-dependent weights by equal weights gives *Group-Balanced Error Rate* (GBER):

$$\mathcal{R}_{\mathrm{GBER}}(h) = \frac{1}{KG} \sum_{y \in \mathcal{Y}} \sum_{a \in \mathcal{A}} \mathbb{P}\big[h(X, A) \neq y \mid Y = y,\ A = a\big]. \tag{1}$$

GBER separates the importance of recognizing a cell from its population frequency: it averages class-conditional errors equally within each group, then averages equally across groups. Correct recognition remains the objective, so improvements in rare cells count as much as equal improvements in prevalent cells.

GBER follows the principle of equally weighting recognition errors across group–class combinations, without directly optimizing a particular fairness metric. Theoretically, GBER also provides a sufficient link to equal opportunity: for each class, a sufficiently small average recognition error across groups limits the corresponding recall disparity. This relation supports the fairness motivation of the equal-weight recognition objective. Detailed analysis is provided in Appendix A.2.

3.3 PRIOR CORRECTION UNDER GBER

GBER gives each group's recognition performance on each class equal importance. We now translate this objective into a decision rule, identifying the component of posterior scores that must be corrected to remove the influence of group-conditional class frequencies.

Theorem 1 (Prior correction)**.** *Under the standing assumptions in Appendix A, a classifier minimizing $\mathcal{R}_{\mathrm{GBER}}$ over all measurable decision rules is*

$$h^*(x, a) = \arg\max_{y \in \mathcal{Y}} \big[\log p(y \mid x, a) - \log p(y \mid a)\big], \tag{2}$$

where $p(y \mid x, a)$ and $p(y \mid a)$ are the true posterior and group-conditional class probabilities, respectively.

Proof sketch. The equal cell weights in GBER give

$$\mathcal{R}_{\mathrm{GBER}}(h) = 1 - \frac{1}{KG} \int_{\mathcal{X}} \sum_{a \in \mathcal{A}} p(x \mid h(x,a), a) \, \mathrm{d}\mu(x).$$

Thus, for each observed (x, a), minimizing GBER selects the class with the greatest group-conditional likelihood:

$$h^*(x,a) = \arg\max_{y \in \mathcal{Y}} p(x \mid y, a). \tag{3}$$

To express this decision in terms of posterior probabilities, Bayes' rule gives

$$\log p(y \mid x, a) = \log p(x \mid y, a) + \log p(y \mid a) - \log p(x \mid a).$$

The posterior therefore includes a group-conditional prior contribution in addition to the class-conditional evidence. Subtracting $\log p(y \mid a)$ recovers the likelihood ordering, since $\log p(x \mid a)$ is constant across candidate classes. This yields Eq. (2); the complete proof is in Appendix A. □

The correction follows from the recognition objective: GBER assigns equal importance to cells whose different frequencies give them unequal prior contributions in posterior scores. Removing this contribution eliminates the frequency advantage in score comparison, giving lower-prior classes a larger relative correction. The resulting rule retains the simple additive form of logit adjustment (Menon et al., 2020), with group-conditional priors determined by the equal-weight recognition objective. This provides the basis for applying the same correction to learned logits.

3.4 Fair logit adjustment

We propose *Fair Logit Adjustment* (FLA), a post-processing method that implements the GBER-derived prior correction on a trained classifier's logits. To accommodate the scale of these learned scores, an adjustment weight $\tau \geq 0$ controls the strength of the prior correction, giving the prediction rule:

$$h_{\mathrm{FLA}}(x,a) \;=\; \arg\max_{y \in \mathcal{Y}} \; \big[f_y(x,a) - \tau \log \hat{\pi}_{y|a} \big]. \tag{4}$$

Here $\hat{\pi}_{y|a}$ estimates $p(y \mid a)$ from a held-out post-processing sample, and τ controls the correction strength relative to the learned-logit scale. For $\tau > 0$, dividing all scores by τ gives the equivalent rule $\arg\max_y [f_y(x,a)/\tau - \log \hat{\pi}_{y|a}]$, exposing the relation between adjustment strength and logit scale. With oracle posterior scores and exact priors, $\tau = 1$ recovers Theorem 1; $\tau = 0$ retains the original predictions. We select a single $\tau \in [0, 2]$ on validation data using the protocol in §4.1. No separate probability-calibration step is needed to apply the rule (Appendix D); its population risk interpretation is detailed in Appendix A.

Prior estimation. For a fixed τ, write the empirical and population offsets as

$$\hat{\Delta}_{y,a} \;=\; \tau \cdot \log \hat{\pi}_{y|a}, \qquad \Delta^*_{y,a} \;=\; \tau \cdot \log \pi_{y|a}. \tag{5}$$

We use maximum likelihood estimates on an estimation sample disjoint from training. Proposition 1 quantifies how errors in these estimates propagate to the offsets. The GBER derivation specifies the conditional-prior target, denoted `prior` in the experiments. We also evaluate the `balance` and `uniform` alternatives as diagnostics of the choice of offset (Table 3). In the binary single-score implementation used for those alternatives, group-frequency terms change the relative class score; this differs from adding a class-independent constant to all logits, which leaves the argmax unchanged.

Proposition 1 (Finite-sample convergence)**.** *Let the post-processing set contain n samples, with $n_a = \sum_y n_{y,a}$ samples from group a (deterministic stratified counts). Assume that, conditional on $A = a$, the labels $Y_1, \ldots, Y_{n_a}$ are i.i.d. with $\mathbb{P}(Y = y \mid A = a) = \pi_{y|a}$ (labels are i.i.d. within each group). Assume $n_{\min}\pi^2_{\min} \geq 2\log(2KG/\delta)$, where $n_{\min} = \min_a n_a$ and $\pi_{\min} = \min_{y,a} \pi_{y|a}$. For any $\delta \in (0,1)$, with probability $\geq 1 - \delta$,*

$$\max_{y,a} \big| \hat{\Delta}_{y,a} - \Delta^*_{y,a} \big| \;\leq\; \frac{2\tau}{\pi_{\min}} \cdot \sqrt{\frac{\log(2KG/\delta)}{2\, n_{\min}}}, \tag{6}$$

where $\tau > 0$, $\pi_{\min} = \min_{y,a} \pi_{y|a}$, and $\Delta^*_{y,a} = \tau \log \pi_{y|a}$ is the exact-prior adjustment. (For $\tau = 0$, $\hat{\Delta}_{y,a} = \Delta^*_{y,a} = 0$ identically and the bound holds trivially.)

Proof. By Hoeffding's inequality and a union bound over all KG cells, $|\hat{\pi}_{y|a} - \pi_{y|a}| \le \sigma := \sqrt{\log(2KG/\delta)/(2n_{\min})}$ simultaneously for all (y, a) with probability $\ge 1 - \delta$. Under the precondition, $\sigma \le \pi_{\min}/2$, so $\hat{\pi}_{y|a} \ge \pi_{\min}/2 > 0$ and $\log \hat{\pi}_{y|a}$ is well-defined. By the mean value theorem applied to $\log$ on $[\pi_{\min}/2, 1]$, $|\log \hat{\pi}_{y|a} - \log \pi_{y|a}| \le |\hat{\pi}_{y|a} - \pi_{y|a}|/(\pi_{\min}/2) = 2\sigma/\pi_{\min}$. Multiplying by τ gives (6). □

Prior estimation becomes less reliable for sparsely represented group–class combinations; Appendix K examines its empirical robustness.

Proposition 1 yields two results on the stability of the empirical correction (Appendix A.1): a GBER excess-risk bound under estimated priors (Theorem 2) and a score-perturbation robustness bound (Proposition 2). Both take the margin-event form $\Pr_{X|Y=y,A=a}(\gamma^*(X, a) \le \text{perturbation})$, where $\gamma^*(x, a)$ denotes the optimal decision margin; excess GBER is thus controlled by the probability mass on examples near the decision boundary. The two bounds unify as the condition $\gamma^*(X, a) \le \eta + 2\varepsilon$ (Corollary 2), and a global logit rescaling $f' = \alpha f$ is absorbed by $\tau' = \alpha\tau$ (Corollary 1).

3.5 In-processing prior correction

Our main theoretical and empirical focus is post-processing. We also describe an in-processing prior-correction variant for settings where retraining is available. Whereas the post-hoc empirical rule (4) shifts the inference logits by $-\tau \log \pi_{y|a}$, the in-processing variant injects a positive prior offset into the pre-softmax activations during training:

$$\mathcal{L}_{\text{FLA}}(x, y, a) \;=\; -\log \frac{\exp\big(f_y(x, a) + \tau \log \pi_{y|a}\big)}{\sum_{k=1}^{K} \exp\big(f_k(x, a) + \tau \log \pi_{k|a}\big)}. \tag{7}$$

This form has a parallel population interpretation to the post-hoc rule. Let $z_y(x, a) = f_y(x, a) + \tau \log \pi_{y|a}$ denote the logits seen by the softmax loss. For the population softmax cross-entropy, minimization over measurable score functions admits minimizers whose shifted scores satisfy

$$z^*_y(x, a) = \log \mathbb{P}(Y = y \mid X = x, A = a) + C(x, a).$$

Equivalently,

$$f^*_y(x, a) = \log \mathbb{P}(Y = y \mid X = x, A = a) - \tau \log \pi_{y|a} + C(x, a).$$

Thus the standard inference rule $\arg\max_y f_y(x, a)$ has the corresponding prior-adjusted population form, without a separate post-hoc intervention. The in-processing variant uses the same empirical prior estimates and adjustment weight as the post-processing rule. This population-level argument identifies the target score shift; it is not a finite-sample optimization guarantee, and its behavior depends on the model class and optimizer. The probability-calibration diagnostic—including group-conditional ECE and a temperature-scaling ablation—is provided in Table 4; the controlled in-processing analysis is reported in §4.4.

4 Experiments

4.1 Setup

We evaluate Adult (Asuncion et al., 2007), COMPAS (Angwin et al., 2022), ACSIncome-2, ACSIncome-5 (Ding et al., 2021), and BiasBios (De-Arteaga et al., 2019) (Communities is a sparse-cell stress test; Appendix J). Tabular predictors are LR, GBDT, and MLP; BiasBios uses BERT-base (Devlin et al., 2019). Baselines are ERM, Hardt-EOpp (binary randomized-threshold post-processing) (Hardt et al., 2016), FairProjection (Alghamdi et al., 2022), and LinearPost (LP) (Xian & Zhao, 2024). LP is the more general successor to FC (Xian et al., 2023), extending its statistical-parity formulation to a unified framework supporting SP, EOpp, and EO. We use LP-EOpp in the main comparison; LP-SP and LP-EO provide metric-specific diagnostics in Appendix G. Across 5 seeds

Table 1: Main results (test means over 5 seeds). Columns report macro F1 and ΔEOpp; bold/underline mark best/second-best. Hardt-EOpp (Hardt et al. 2016), a binary randomized-threshold post-processor, is binary-only. FLA, LP-EOpp, and FairProjection use validation-selected operating points under the same two-percentage-point accuracy tolerance.

	Base		Hardt-EOpp		FairProj.		LP-EOpp		FLA	
	F1	EOpp	F1	EOpp	F1	EOpp	F1	EOpp	F1	EOpp
Adult/LR	<u>0.782</u>	0.102	0.751	0.035	0.781	0.023	0.779	**0.018**	**0.791**	<u>0.019</u>
Adult/GBDT	<u>0.812</u>	0.105	0.783	0.036	0.812	0.024	0.811	<u>0.017</u>	**0.813**	**0.013**
COMPAS/LR	**0.682**	0.224	0.666	0.034	<u>0.681</u>	0.035	0.679	**0.029**	0.674	<u>0.031</u>
COMPAS/GBDT	**0.658**	0.169	0.629	0.050	0.645	<u>0.029</u>	<u>0.647</u>	0.047	0.644	**0.013**
ACS-2/LR	**0.757**	0.182	0.751	0.006	0.756	<u>0.005</u>	0.754	0.005	**0.757**	**0.002**
ACS-2/GBDT	**0.798**	0.120	0.786	0.003	<u>0.795</u>	0.003	0.794	<u>0.002</u>	0.792	**0.001**
ACS-2/MLP	**0.790**	0.119	0.777	0.004	<u>0.787</u>	0.005	0.786	<u>0.003</u>	0.784	**0.001**
ACS-5/LR	**0.456**	0.488	—	—	<u>0.454</u>	0.203	0.453	**0.088**	<u>0.454</u>	<u>0.164</u>
ACS-5/GBDT	**0.523**	0.399	—	—	<u>0.521</u>	0.144	0.521	**0.094**	0.520	<u>0.140</u>
ACS-5/MLP	**0.514**	0.485	—	—	<u>0.513</u>	0.223	0.512	**0.102**	0.512	<u>0.190</u>
BiasBios/BERT	**0.816**	0.417	—	—	0.796	**0.187**	<u>0.814</u>	0.288	0.771	<u>0.265</u>

(33–37), all post-processors are validation-selected on ΔEOpp under a two-percentage-point accuracy tolerance relative to the base model. Table 1 reports macro F1 and ΔEOpp; companion accuracy, ΔSP, and ΔEO are in Appendix F. FLA estimates $\hat{\pi}_{y|a}$ from label/group counts on the post-processing split and sweeps $\tau \in [0, 2]$ over 2001 values.

4.2 Main results

Tabular benchmark results. Across the ten tabular dataset–model pairs, FLA is always best or second-best on ΔEOpp in Table 1. The binary tabular settings show the strongest version of the result: on Adult, COMPAS, and ACSIncome-2, a single deterministic prior-offset sweep often matches or outperforms optimization-based post-processing under the same accuracy guard. Additional comparisons with FRAPPÉ (Tifrea et al., 2023) and MBS (Chen et al., 2024) are reported in Appendix E.2, as the implementations used here support binary classification only.

ACSIncome-5 marks the harder multi-class tabular boundary. With 25 (y, a) cells, LP-EOpp optimizes its correction under an EOpp constraint and obtains lower selected ΔEOpp than FLA on all three predictors. FLA is second-best on ΔEOpp, keeps macro F1-Score within a few thousandths of the base model, and improves the group-recall structure across all five income classes, as shown in Appendix C. The tabular evidence thus supports the group-prior correction as a competitive closed-form approach when cell counts are reliable, while ACSIncome-5 illustrates a setting where LP-EOpp achieves lower disparity.

Text benchmark results. BiasBios separates the scaling and utility trade-offs from the tabular story. Here $K{=}28$ and $KG{=}56$, many occupations have strong gender-conditional priors, and LP post-processing becomes expensive. FLA reduces ΔEOpp from the base model's 0.417 to 0.265, slightly improving over LP-EOpp's 0.288, while reducing construction time by orders of magnitude (§4.3). The utility picture is more nuanced: the selected FLA point satisfies the two-percentage-point accuracy guard but sacrifices more macro F1-Score than LP-EOpp. FairProjection achieves the lowest disparity in this row at the cost of substantially higher prediction-time interpolation overhead (Figure 2); FLA is the only method in this row combining competitive disparity with negligible prediction overhead. The interpretation is consistent with the mechanism: FLA and LP-EOpp have highly correlated ΔEOpp improvement patterns (Table 13, Appendix H), showing similar variation in their improvements across the evaluated settings. On a 28-way semantic task, aggressive prior offsets can also move borderline examples among rare occupations; §5 discusses when projection or LP methods remain preferable.

Even without tuning the adjustment strength, FLA with $\tau = 1$ reduces mean ΔEOpp relative to the base model in all 11 main settings; Appendix E.4 reports the corresponding utility trade-offs.

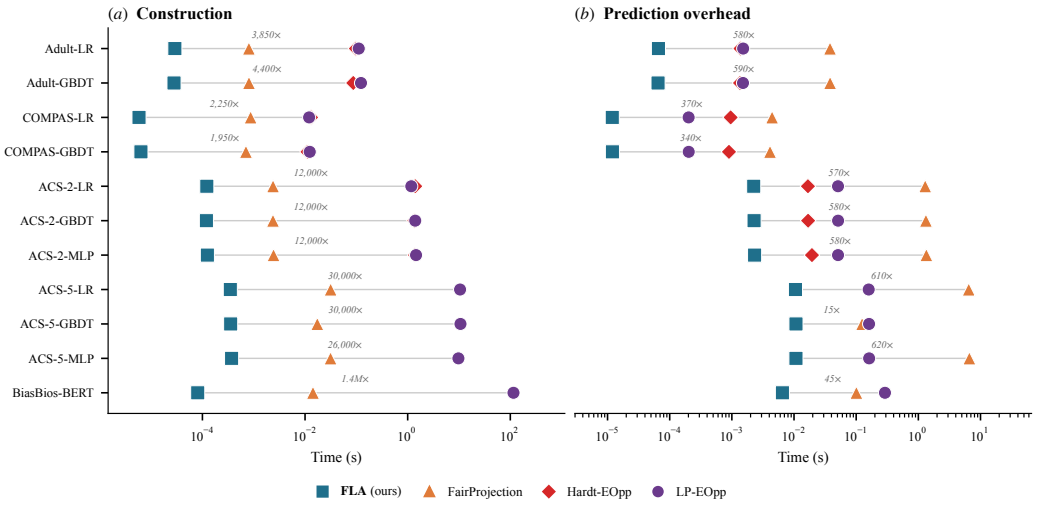

Figure 2: Post-processing construction time and prediction overhead (log scale, mean over 5 seeds) for all baselines. Panel (a) reports construction time; panel (b) reports prediction overhead after fitting. In construction, LP-EOpp incurs substantial fitting cost on the larger benchmarks. Connectors show the LP-EOpp / FLA construction-speed ratio; Hardt-EOpp (binary-only) is omitted for ACSIncome-5 and BiasBios.

Cross-metric diagnostic. At the same EOpp-selected operating points, FLA also reduces ΔEO relative to ERM in all 11 dataset–model settings (median absolute drop 0.095); full cross-metric diagnostics for the main selection are in Tables 9–10, and EO/SP-selected diagnostics are in Tables 11–12.

Interpretability. Unlike post-processors that remap the output distribution, FLA's prior-correction view yields an interpretable intervention: fairness correction is formulated as the removal of a group-conditional prior term from the logits. The resulting operation is an explicit $K \times G$ table of group-class offsets. For group a, class y receives the deterministic offset $-\tau \log \hat{\pi}_{y|a}$, which increases the relative score of low-prior group-class cells and removes the prevalence advantage of high-prior cells internalized in the ERM logits. A prediction changes only when this correction exceeds the original logit margin between candidate labels, making each changed decision attributable to an explicit comparison between model evidence and the group-conditional prior correction. Figure 1 illustrates the recall-level effect on BiasBios (both bias directions); the multi-group ACSIncome-5 analogue (25 group-class cells) is in Appendix C. Non-prior target diagnostics are in Appendix B.

4.3 Computational efficiency

Figure 2 shows construction time and prediction overhead for all post-processors on shared log scales. FLA and FairProjection have sub-second construction times across our benchmarks; Hardt-EOpp's ACSIncome-2 timing includes the `fairlearn` threshold search. LP-EOpp fits its fairness correction by solving an empirical linear program, with construction time reaching 115 s on BiasBios. FLA is the fastest: 1.9×10^3–$1.4\times10^6\times$ faster than LP-EOpp and approximately $100\times$ faster than FairProjection in recorded construction time. The prediction-overhead panel shows the same deployment pattern: FLA remains a negligible logit-offset computation, while FairProjection's interpolation overhead becomes visible on the larger tabular and text settings. LP-EOpp did not solve successfully in some settings with tight fairness tolerances on BiasBios and ACSIncome-5. FLA applies its prior offset without an optimization solver.

Table 2 summarizes the deployment profile.

Table 2: Deployment profile. Randomness denotes prediction-time stochasticity.

Method	Train	PP Fit	Inference	Randomness	Code	Guarantee
Adversarial training	Hours	—	High	None	Complex	No
Hardt-EOpp	Zero	Seconds	Low	Prediction	Simple	Yes (EOpp, binary)
FairProjection	Zero	Seconds	Medium	None	Complex	Yes (projection)
LP-EOpp	Zero	Min–Hours	Medium	Small noise	Complex	Yes (EOpp)
FLA (ours)	**Zero**	**No fitting***	**Negligible**	**None**	**1 line**	**Yes (GBER)**

Note. LP denotes LinearPost (Xian & Zhao, 2024); its fairness guarantee holds under the predictor assumptions of that work, and the implementation adds small random perturbations to prediction risks. *For FLA, this column reports empirical prior estimation and offset construction (under 0.1 s in all our settings), not optimization or model fitting.

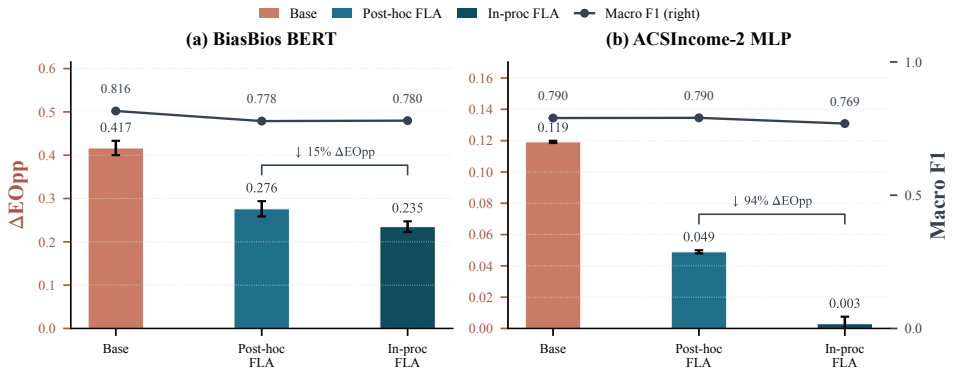

Figure 3: In-processing prior correction at matched $\tau = 1$. Bars show ΔEOpp; the dark curve shows macro F1. Brackets show the change induced by applying the same prior offset during training.

4.4 In-processing analysis

Figure 3 evaluates the in-processing extension of the same prior-correction principle at matched $\tau = 1$. On BiasBios BERT, it reduces ΔEOpp from 0.276 to 0.235 with nearly unchanged utility. On ACSIncome-2 MLP, it reduces the EOpp gap from 0.049 to 0.003 but lowers accuracy and macro F1. These results illustrate how the principle can also be applied during training, with different fairness–utility trade-offs across the two settings.

5 Discussion and Limitations

Method selection. FLA provides a deterministic, solver-free correction with low construction and prediction cost (§4.3). Its use of empirical group-conditional priors makes the support of individual group–class cells an important consideration. In the sparse Communities setting, validation selects an adjustment close to zero, while LP-EOpp achieves lower disparity (Table 14). The subsampling analysis in Appendix K further examines sensitivity to the amount of data available for prior estimation.

Deployment scope. Post-hoc FLA follows the standard group-aware post-processing setting: like Hardt (Hardt et al., 2016) and the LP setting evaluated here (Xian & Zhao, 2024), it uses the sensitive attribute at inference time. When this is unavailable, the in-processing variant provides the same prior-correction principle using the attribute during training rather than at prediction time.

6 Conclusion

We view group unfairness through the lens of long-tail learning: a class that is common overall may be underrepresented within a particular group, and the resulting prior imbalance can be absorbed into

learned logits, contributing to unequal recognition. This perspective leads us to GBER, which gives equal importance to recognition errors across group–class combinations. From its Bayes rule, we derive Fair Logit Adjustment (FLA), translating this principle into a lightweight, deterministic, and interpretable correction of model outputs without retraining. Across multiple benchmarks and model families, we demonstrate that FLA achieves competitive fairness–utility trade-offs compared with state-of-the-art post-processing methods, with minimal computational overhead. Our work connects a long-tail view of unequal recognition to a practical post-processing approach, showing how correcting absorbed group-conditional priors can improve fairness in existing classifiers.

Reproducibility Statement

The supplementary code provides an anonymous, self-contained implementation of post-hoc FLA and a runnable demo of prior estimation and logit correction. Section 3 specifies the method, and Appendix A states the assumptions and gives the full proofs. Experimental settings, hyperparameters, and data and software sources are provided in Section 4.1, Appendix I, and Appendix M; additional evaluation protocols and results are reported in Appendix E.

AI Use Statement

AI tools were used to assist with manuscript editing, checking mathematical derivations, and refining proofs and their exposition. The authors reviewed the resulting changes and take full responsibility for all claims and proofs.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pp. 60–69. PMLR, 2018.

Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, Peter Michalak, Shahab Asoodeh, and Flavio Calmon. Beyond adult and compas: Fair multi-class prediction via information projection. *Advances in Neural Information Processing Systems*, 35:38747–38760, 2022.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pp. 254–264. Auerbach Publications, 2022.

Arthur Asuncion, David Newman, et al. Uci machine learning repository, 2007.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Wenlong Chen, Yegor Klochkov, and Yang Liu. Post-hoc bias scoring is optimal for fair classification. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 32825–32852, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/8b76bc305ccfb95010081d12b85c3b24-Paper-Conference.pdf.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 120–128, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Federico Di Gennaro, Thibault Laugel, Vincent Grari, Xavier Renard, and Marcin Detyniecki. Post-processing fairness with minimal changes. *arXiv preprint arXiv:2408.15096*, 2024.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Naoya Hasegawa and Issei Sato. Multiplicative logit adjustment approximates neural-collapse-aware decision boundary adjustment. *arXiv preprint arXiv:2409.17582*, 2024.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1gRTCVFvB.

Lydia T. Liu, Max Simchowitz, and Moritz Hardt. The implicit fairness criterion of unconstrained learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4051–4060. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/liu19f.html.

Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, pp. 107–118. PMLR, 2018.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

Jiawei Ren, Cunjun Yu, shunan sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4175–4186. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/2ba61cc3a8f44143e1f2f13b2b729ab3-Paper.pdf.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8346–8356. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/sagawa20a.html.

Alexandru Tifrea, Preethi Lahoti, Ben Packer, Yoni Halpern, Ahmad Beirami, and Flavien Prost. Frappé: A group fairness framework for post-processing everything. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:265659394.

Ruicheng Xian and Han Zhao. A unified post-processing framework for group fairness in classification, 2024. URL https://arxiv.org/abs/2405.04025.

Ruicheng Xian, Lang Yin, and Han Zhao. Fair and optimal classification via post-processing. In *International conference on machine learning*, pp. 37977–38012. PMLR, 2023.

Zhengzhuo Xu, Zenghao Chai, and Chun Yuan. Towards calibrated model for long-tailed visual recognition from prior perspective. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 7139–7152. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/39ae2ed11b14a4ccb41d35e9d1ba5d11-Paper.pdf.

Xianli Zeng, Kevin Jiang, Guang Cheng, and Edgar Dobriban. Bayes-optimal fair classification with linear disparity constraints via pre-, in-, and post-processing. *Journal of Machine Learning Research*, 27(65):1–87, 2026.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Xiaoling Zhou, Ou Wu, and Nan Yang. Class and attribute-aware logit adjustment for generalized long-tail learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22991–22999, 2025.

A PROOFS AND STABILITY THEOREMS

Standing assumptions. The following technical conditions are assumed throughout Section 3 and are referenced as needed in the proofs below: (i) *Positivity*: $\pi_{y,a} > 0$ for all (y, a), so conditional expectations and $\log \pi_{y|a}$ are well-defined; (ii) *Common dominating measure*: (X, Y, A) admits a joint density with respect to a product measure $\mu \otimes \#$, so Tonelli's theorem applies to the GBER integral; (iii) *Measurable selection*: a measurable selection of the pointwise GBER minimizer exists, which is guaranteed when $\mathcal{X}$ is a standard Borel space; (iv) *Tie-breaking*: ties in the argmax occur on a μ-null set and may be broken arbitrarily without affecting GBER.

We give a self-contained proof in two steps.

STEP 1: POINT-WISE REDUCTION OF GBER MINIMIZATION

Lemma 1 (Point-wise reduction)**.** *Let $h : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ be any measurable classifier. For any fixed $(x_0, a_0) \in \mathcal{X} \times \mathcal{A}$, the GBER-optimal prediction satisfies:*

$$h^*(x_0, a_0) \;=\; \arg\max_{y \in \mathcal{Y}} \; p_{X|Y,A}(x_0 \mid y, a_0). \tag{8}$$

Proof. Write GBER explicitly:

$$\begin{aligned}\mathcal{R}_{\mathrm{GBER}}(h) &= \frac{1}{KG} \sum_{y \in \mathcal{Y}} \sum_{a \in \mathcal{A}} \mathbb{E}_{X|Y=y,A=a}\big[\mathbf{1}(h(X, A) \neq y)\big] \\ &= \frac{1}{KG} \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} \sum_{a \in \mathcal{A}} \mathbf{1}(h(x, a) \neq y)\, p_{X|Y,A}(x \mid y, a)\, \mathrm{d}\mu(x). \end{aligned} \tag{9}$$

The integrand can be minimized point-wise over x. For fixed (x_0, a_0) and variable $\hat{y} = h(x_0, a_0)$, the relevant contribution to the sum over a is (only the $a = a_0$ term contains $h(x_0, a_0)$):

$$\sum_{y \in \mathcal{Y}} \mathbf{1}(\hat{y} \neq y)\, p_{X|Y,A}(x_0 \mid y, a_0) \;=\; \sum_{y} p_{X|Y,A}(x_0 \mid y, a_0) - p_{X|Y,A}(x_0 \mid \hat{y}, a_0). \tag{10}$$

Since $\sum_y p_{X|Y,A}(x_0 \mid y, a_0)$ does not depend on $\hat{y}$, minimizing the above is equivalent to maximizing $p_{X|Y,A}(x_0 \mid \hat{y}, a_0)$ over $\hat{y}$, giving (8). □

STEP 2: POSTERIOR-LOGIT FORM

Let $s^*_y(x, a) = \log p(y \mid x, a)$ denote the oracle posterior logit, and recall that $\pi_{y|a} = p(y \mid a)$. By Bayes' rule,

$$s^*_y(x, a) \;=\; \log p_{X|Y,A}(x \mid y, a) \;+\; \log \pi_{y|a} \;-\; \log p_{X|A}(x \mid a). \tag{11}$$

Rearranging gives

$$\log p_{X|Y,A}(x \mid y, a) \;=\; s^*_y(x, a) \;-\; \log \pi_{y|a} \;+\; \log p_{X|A}(x \mid a). \tag{12}$$

Since $\log p_{X|A}(x \mid a)$ is constant over y, the argmax in (8) is:

$$h^*(x, a) \;=\; \arg\max_{y} \; p_{X|Y,A}(x \mid y, a) \;=\; \arg\max_{y} \; \big[s^*_y(x, a) - \log \pi_{y|a}\big], \tag{13}$$

which is Eq. (2). □

Equivalently, one may subtract $\log p(y, a)$, since $\log p(y, a) = \log p(y \mid a) + \log p(a)$ and the last term is constant across candidate classes for a fixed a.

THE R_τ FAMILY

At the population level, $\arg\max_y [s^*_y(x, a) - \tau \log \pi_{y|a}]$ minimizes

$$R_\tau(h) = \frac{1}{KG} \sum_{y,a} \pi_{y|a}^{1-\tau} \mathbb{P}(h(X, A) \neq y \mid Y = y, A = a).$$

At $\tau = 0$, R_0 equals the group-uniform average of per-group error rates:

$$R_0 = \frac{1}{K} \cdot \frac{1}{G} \sum_a \mathbb{P}(h \neq Y \mid A{=}a),$$

which differs from the natural population error $\sum_a \mathbb{P}(A{=}a)\mathbb{P}(h \neq Y \mid A{=}a)$ unless A is uniform. Its minimizer is the standard posterior classifier $h^*(x, a) = \arg\max_y \mathbb{P}(Y = y \mid X = x, A = a)$; at $\tau = 1$, $R_1 = \mathcal{R}_{\mathrm{GBER}}$.

A.1 Stability under prior estimation and score perturbations

Theorem 2 (GBER stability under estimated priors)**.** *Let $\tau \geq 0$, let $\hat{h}$ denote* FLA *with empirically estimated priors $\hat{\pi}_{y|a}$, and let h^*_τ denote the exact-prior empirical classifier at the same τ and with the same scores f (using $\pi_{y|a}$ instead of $\hat{\pi}_{y|a}$). If $f = s^*$ and $\tau = 1$, then h^*_1 is the GBER-Bayes-optimal rule (Theorem 1), so the bound specializes to a true excess-GBER bound over the global minimum. Define the* optimal decision margin *at (x, a) as*

$$\gamma^*_\tau(x, a) \;=\; \big[f_{y^*}(x, a) - \tau \log \pi_{y^*|a}\big] \;-\; \max_{y \neq y^*} \big[f_y(x, a) - \tau \log \pi_{y|a}\big], \tag{14}$$

where $y^* = h^*_\tau(x, a)$. We write γ^* for $\gamma^*_{\tau=1}$ in Proposition 2. With probability $\geq 1 - \delta$ over the postprocessing set,

$$\mathcal{R}_{\mathrm{GBER}}(\hat{h}) - \mathcal{R}_{\mathrm{GBER}}(h^*_\tau) \;\leq\; \frac{1}{KG} \sum_{y,a} \mathbb{P}_{X|Y=y,A=a}\big(\gamma^*_\tau(X, a) \leq 2\varepsilon\big), \tag{15}$$

For $\tau > 1$ the reference rule h^*_τ is no longer the GBER-Bayes-optimal rule (Theorem 1) but the cost-sensitive R_τ minimizer of Appendix A; the inequality and its proof are unchanged. Here

$$\varepsilon \;=\; \frac{2\tau}{\pi_{\min}} \sqrt{\frac{\log(2KG/\delta)}{2\, n_{\min}}}$$

is the bound from Proposition 1 (right-hand side of (6)*).*

Proof of Theorem 2 (GBER Stability)

Proof. By Proposition 1, with probability $\geq 1 - \delta$: $|\hat{\Delta}_{y,a} - \Delta^*_{y,a}| \leq \varepsilon$ for all (y, a) simultaneously.

On the good event $\hat{\pi}_{y|a} > 0$ for all (y, a) (guaranteed by Proposition 1's precondition $n_{\min}\pi^2_{\min} \geq 2\log(2KG/\delta)$), $\hat{\Delta}_{y,a} = \tau \log \hat{\pi}_{y|a}$ is well-defined and $\hat{h}$ is the argmax classifier using $\hat{\Delta}$; on the bad event (probability $< \delta$) define $\hat{h}$ arbitrarily.

Fix any (x, a) with $y^* = h^*_\tau(x, a)$. If $\hat{h}(x, a) \neq h^*_\tau(x, a)$, there exists $y \neq y^*$ such that

$$\big[f_{y^*}(x, a) - \hat{\Delta}_{y^*,a}\big] \leq \big[f_y(x, a) - \hat{\Delta}_{y,a}\big]. \tag{16}$$

Substituting $\hat{\Delta} = \Delta^* + (\hat{\Delta} - \Delta^*)$ and using the definition of $\gamma^*_\tau(x, a)$, condition (16) requires

$$(\hat{\Delta}_{y^*} - \Delta^*_{y^*}) - (\hat{\Delta}_y - \Delta^*_y) \geq \gamma^*_\tau(x, a). \tag{17}$$

Since $|\hat{\Delta}_y - \Delta^*_y| \leq \varepsilon$ for all y, the left-hand side is at most 2ε. Therefore $\hat{h}(x, a) \neq h^*_\tau(x, a)$ *implies* $\gamma^*_\tau(x, a) \leq 2\varepsilon$.

Summing over cells:

$$\begin{aligned}
\mathcal{R}_{\mathrm{GBER}}(\hat{h}) - \mathcal{R}_{\mathrm{GBER}}(h^*_\tau) &= \frac{1}{KG} \sum_{y,a} \mathbb{E}\big[\mathbf{1}\{\hat{h}(X, a) \neq y\} - \mathbf{1}\{h^*_\tau(X, a) \neq y\} \mid Y{=}y, A{=}a\big] \\
&\leq \frac{1}{KG} \sum_{y,a} \mathbb{P}_{X|Y=y,A=a}\big(\hat{h}(X, a) \neq h^*_\tau(X, a)\big) \\
&\leq \frac{1}{KG} \sum_{y,a} \mathbb{P}_{X|Y=y,A=a}\big(\gamma^*_\tau(X, a) \leq 2\varepsilon\big).
\end{aligned} \tag{18}$$

The first inequality holds because if $\hat{h} = h^*_\tau$ the indicator difference is zero; if they differ, it is at most $+1$. □

Proposition 2 (Robustness of the empirical rule to score perturbations)**.** *Assume exact priors ($\hat{\pi}_{y|a} = \pi_{y|a}$) and $\tau = 1$. Let h^* be the GBER-Bayes-optimal rule from Theorem 1, based on oracle posterior scores s^*. Let $\gamma^*(x, a)$ be its adjusted-score margin,*

$$\gamma^*(x,a) = \big[s^*_{y^*}(x,a) - \log \pi_{y^*|a}\big] - \max_{y \neq y^*} \big[s^*_y(x,a) - \log \pi_{y|a}\big], \quad y^* = h^*(x,a).$$

Suppose the empirical score satisfies $f_y(x,a) = \log \mathbb{P}(Y = y \mid X = x, A = a) + C(x,a) + \epsilon_y(x,a)$, where C is y-independent and ϵ_y is a class-dependent score perturbation. Let $\eta = \sup_{x,a} \big(\max_y \epsilon_y(x,a) - \min_y \epsilon_y(x,a)\big)$ denote the worst-case inter-class spread. Let h_η denote FLA *with $\tau = 1$ applied to the perturbed scores (using exact priors $\pi_{y|a}$). Then*

$$\mathcal{R}_{\mathrm{GBER}}(h_\eta) - \mathcal{R}_{\mathrm{GBER}}(h^*) \;\le\; \frac{1}{KG} \sum_{y,a} \mathbb{P}_{X|Y=y,A=a}\big(\gamma^*(X,a) \le \eta\big), \tag{19}$$

The bound is tightest when the inter-class score spread η is small relative to the typical decision margin. A special case is given in the Corollary below.

PROOF OF PROPOSITION 2 (SCORE-PERTURBATION ROBUSTNESS)

Proof. Here $\gamma^*(x,a)$ denotes the ideal decision margin computed from $f^0_y = \log \mathbb{P}(Y = y \mid X = x, A = a) + C(x,a)$ ($\epsilon_y \equiv 0$) at $\tau = 1$. Under the stated model, applying FLA gives adjusted score

$$\tilde{f}_y(x,a) = \log \mathbb{P}(Y = y \mid X = x, A = a) - \log \pi_{y|a} + C(x,a) + \epsilon_y(x,a). \tag{20}$$

The optimal classifier uses $\epsilon_y \equiv 0$. Therefore $h_\eta(x,a) \neq h^*(x,a)$ only if there exists $y \neq y^*$ with

$$\epsilon_y(x,a) - \epsilon_{y^*}(x,a) \ge \gamma^*(x,a). \tag{21}$$

Since $\epsilon_y - \epsilon_{y^*} \le \eta$ by the definition of η, the error event implies $\gamma^*(x,a) \le \eta$. Summing as in the previous proof gives (19). □

Corollary 1 (τ absorbs global logit scale)**.** *If $f'_y(x,a) = \alpha \cdot f_y(x,a)$ for some $\alpha > 0$ (e.g. temperature scaling at $T = 1/\alpha$), then* FLA *with $\tau' = \alpha\tau$ applied to f' produces identical decisions to* FLA *with τ applied to f.*

Proof. $\arg\max_y [f'_y - \tau' \log \pi_{y|a}] = \arg\max_y [\alpha f_y - \alpha\tau \log \pi_{y|a}] = \arg\max_y [\alpha(f_y - \tau \log \pi_{y|a})] = \arg\max_y [f_y - \tau \log \pi_{y|a}]$. □

Corollary 2 (Combined robustness)**.** *Assume $\tau = 1$. Under the conditions of Proposition 1 and Proposition 2, with probability $\ge 1 - \delta$,*

$$\mathcal{R}_{\mathrm{GBER}}(\hat{h}_\eta) - \mathcal{R}_{\mathrm{GBER}}(h^*) \;\le\; \frac{1}{KG} \sum_{y,a} \mathbb{P}_{X|Y=y,A=a}\big(\gamma^*(X,a) \le \eta + 2\varepsilon\big), \tag{22}$$

where $\hat{h}_\eta$ applies FLA *with estimated priors $\hat{\pi}$ to perturbed scores (inter-class spread η).*

Proof. Write $\delta_y(x,a) = \Delta^*_y - \hat{\Delta}_y = \log \pi_{y|a} - \log \hat{\pi}_{y|a}$ for the estimation perturbation; by Proposition 1, $|\delta_y| \le \varepsilon$ for all (y,a) simultaneously with probability $\ge 1 - \delta$. The combined adjusted score for class y is $\tilde{f}_y = \log \mathbb{P}(Y = y \mid X = x, A = a) - \log \pi_{y|a} + C + \epsilon_y + \delta_y$. An error at (x,a) requires $\exists\, y \neq y^*$ with $(\epsilon_y + \delta_y) - (\epsilon_{y^*} + \delta_{y^*}) > \gamma^*(x,a)$. By the triangle inequality: $(\epsilon_y - \epsilon_{y^*}) + (\delta_y - \delta_{y^*}) \le |\epsilon_y - \epsilon_{y^*}| + |\delta_y| + |\delta_{y^*}| \le \eta + 2\varepsilon$. Hence error implies $\gamma^*(x,a) \le \eta + 2\varepsilon$, and summing over (y,a) as in Theorem 2 gives (22). □

A.2 GBER AND EQUAL OPPORTUNITY

A sufficient relation through classwise recognition error. Fix a classifier h and a class y. Write $e_{y,a} = 1 - \mathrm{TPR}_{y,a}$ for its conditional error in group a and

$$\mu_y = \frac{1}{G} \sum_a e_{y,a} = \mathcal{R}_{\mathrm{GBER},y}(h)$$

for its classwise GBER. Since recall and error sum to one in each cell, the equal-opportunity disparity is the range of the group errors. Nonnegativity of these errors gives

$$\Delta\mathrm{EOpp}(h, y) = \max_a e_{y,a} - \min_a e_{y,a} \le \max_a e_{y,a} \le \sum_a e_{y,a} = G\mu_y. \tag{23}$$

Thus sufficiently small classwise GBER controls the corresponding recall disparity, with every group's recognition error receiving equal weight regardless of its prevalence. This supplies a theoretical connection between the equal-weight recognition objective and equal opportunity. It is a sufficient relation: a decrease in the mean error need not imply a decrease in its range.

Refinement using variation across groups. For a given mean recognition error, its distribution across groups also matters to the recall disparity. Define

$$\sigma_y^2 = \frac{1}{G}\sum_a (e_{y,a} - \mu_y)^2.$$

For two distinct groups a, a', the inequality $(u - v)^2 \le 2(u^2 + v^2)$ yields

$$\begin{aligned}(e_{y,a} - e_{y,a'})^2 &\le 2\big[(e_{y,a} - \mu_y)^2 + (e_{y,a'} - \mu_y)^2\big] \\ &\le 2\sum_b (e_{y,b} - \mu_y)^2 = 2G\sigma_y^2.\end{aligned}$$

Taking square roots and maximizing over group pairs, then combining with (23), gives

$$\Delta\mathrm{EOpp}(h, y) \le \min\big\{G\mu_y, \sqrt{2G}\,\sigma_y\big\}. \tag{24}$$

The first term uses only the average recognition error; the second uses its variation across groups to refine the bound. For $G = 2$, $\sigma_y = |e_{y,1} - e_{y,2}|/2$, so the second term equals the recall disparity. If all group errors coincide, both the disparity and σ_y are zero; this also covers the single-group case.

B TARGET-DISTRIBUTION DIAGNOSTIC

The theory and main experiments use the conditional-prior correction $\log \pi_{y|a}$. Two alternative targets are superficially appealing: `balance` applies a group-marginal offset, while `uniform` pushes these marginals toward an equal group target. These choices reflect common intuitions about fairness as distributional balancing, but they do not match the posterior-score form of the GBER Bayes rule, which removes the class-dependent conditional-prior term. In the binary single-score convention, the group-marginal offset contributes a threshold shift proportional to $\tau \log \mathbb{P}(A = a)$; this can over-correct by penalizing groups with different base rates, rather than correcting the within-group class prior $\pi_{y|a}$. We therefore treat them as negative-control diagnostics rather than alternative FLA variants.

Table 3 shows the same pattern in multi-class ΔEOpp (lower is better) under EOpp-oriented validation selection.

Table 3: Target-distribution diagnostics (ΔEOpp; lower is better).

Dataset/model	Base	`prior`	`balance`	`uniform`
ACSIncome-5/LR	0.488	**0.164**	0.487	0.488
ACSIncome-5/GBDT	0.399	**0.140**	0.403	0.401
ACSIncome-5/MLP	0.485	**0.190**	0.485	0.485
BiasBios/BERT	0.417	**0.265**	0.417	0.417

The same pattern appears in earlier binary log-reg diagnostics: for Adult/COMPAS/ACSIncome-2, the best EOpp values under `prior` are $0.008/0.035/0.051$, compared with $0.107/0.166/0.167$ for `balance` and $0.108/0.168/0.167$ for `uniform`. Together, these results support the mechanism-specific choice of `prior`: generic prior-like offsets leave EOpp close to the unadjusted model on these benchmarks, whereas the conditional-prior correction provides the substantive reduction reported in the main results.

C MULTI-GROUP TPR VISUALIZATION (ACSINCOME-5)

Figure 4 shows group-conditional TPR radar charts for ACSIncome-5 GBDT (5 income classes $\times$ 5 racial groups, 25 cells). Each polygon traces the mean TPR for one racial group across income classes. ERM produces widely spread polygons, reflecting prior-induced recall disparities. FLA ($\tau = 1$, `prior`) contracts the polygons toward a common shape, reducing mean inter-group TPR standard deviation from 0.064 to 0.027 (-58%) without retraining.

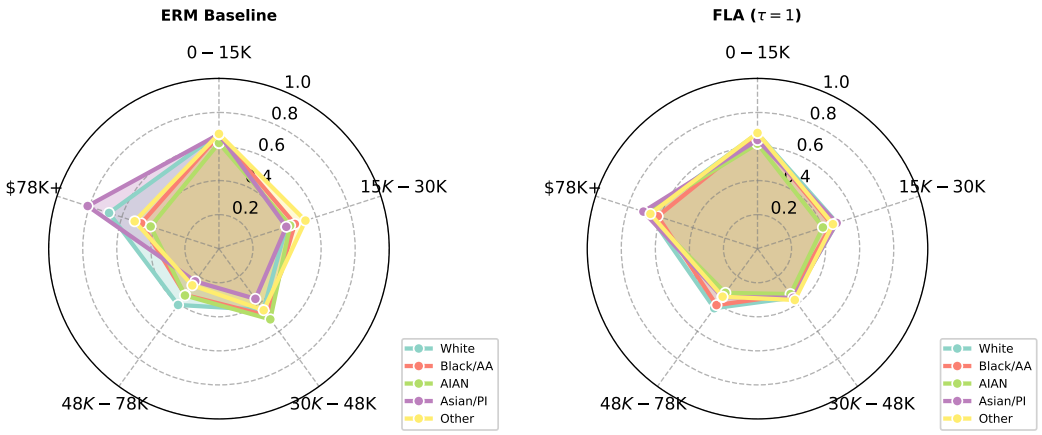

Figure 4: ACSIncome-5 GBDT group-conditional TPR radar charts (5 income classes, 5 racial groups). **Left (ERM)**: wide inter-group spread across income classes. **Right (FLA, $\tau = 1$)**: polygons converge, reducing mean inter-group TPR std from 0.064 to 0.027 (-58%).

D TEMPERATURE SCALING AND PROBABILITY CALIBRATION DIAGNOSTIC

Score scale and calibration. Applying FLA requires no separate probability-calibration step. A global positive logit rescaling can be absorbed by $\tau' = \alpha\tau$ (Corollary 1); class-dependent score perturbations are analyzed in Proposition 2. The following diagnostics examine probability calibration alongside the fairness of the resulting class predictions.

Diagnostic measurements. We report two metrics on the base model and FLA-adjusted output: (i) **Standard ECE** (Guo et al., 2017): top-label expected calibration error (confidence = max softmax, 15 equal-width bins); (ii) **Group-conditional ECE gap**: $\max_a \mathrm{ECE}_a - \min_a \mathrm{ECE}_a$ (inter-group calibration disparity).

Results. Table 4 reports results on Adult/LR and ACSIncome-2/LR averaged over 3 seeds. FLA increases standard ECE relative to the base model (0.064 vs 0.006 on Adult), showing that the correction also changes probability calibration. The reported EOpp and EO disparities are computed from class predictions; ECE provides a separate assessment of the confidence scores.

Temperature scaling ablation. We apply temperature scaling (Guo et al., 2017) to the base model on the post-processing set before FLA ("Base + Temp. Scaling" rows). The reported fairness improvement is unchanged by this preprocessing, consistent with the score-scale relation in Corollary 1.

Table 4: Calibration metrics (3-seed average). ECE: lower = better calibrated. Group ECE gap: $\max_a \text{ECE}_a - \min_a \text{ECE}_a$. ECE assesses probability calibration, while EOpp and EO assess disparities in class predictions.

Dataset	Method	ECE↓	Group ECE gap↓
Adult / LR	Base	0.006	0.004
	FLA (val-tuned τ)	0.064	0.044
	Base + Temp. Scaling	0.007	0.006
	FLA + Temp. Scaling	0.061	0.043
ACS-2 / LR	Base	0.002	0.001
	FLA (val-tuned τ)	0.087	0.036
	Base + Temp. Scaling	0.002	0.002
	FLA + Temp. Scaling	0.087	0.036

E Result reporting and reproducibility notes

Main results report means over five seeds; Table 5 provides standard deviations for Base, LP-EOpp, and FLA. Key reproducibility points: all tabular experiments use seeds 33–37; BiasBios BERT uses the same seed range with 3 training epochs. The anonymous supplementary package isolates the post-hoc FLA implementation and a runnable demo, while Appendix I and Appendix M specify the full evaluation protocol.

E.1 Variability across evaluation seeds

Table 5 reports mean and standard deviation for Base, LP-EOpp, and FLA at the operating points in Table 1. These quantify variation across the five evaluation splits of the frozen predictors, rather than variation from retraining each predictor.

Table 5: Main-result variability over five seeds (mean $\pm$ standard deviation). F1 denotes macro F1; EOpp denotes ΔEOpp.

	Base		LP-EOpp		FLA	
	F1	EOpp	F1	EOpp	F1	EOpp
Adult/LR	0.7823 ± 0.0031	0.1020 ± 0.0121	0.7789 ± 0.0035	0.0179 ± 0.0192	0.7908 ± 0.0030	0.0195 ± 0.0083
Adult/GBDT	0.8123 ± 0.0021	0.1046 ± 0.0095	0.8111 ± 0.0024	0.0170 ± 0.0170	0.8130 ± 0.0012	0.0128 ± 0.0080
COMPAS/LR	0.6818 ± 0.0078	0.2242 ± 0.0230	0.6787 ± 0.0070	0.0295 ± 0.0209	0.6742 ± 0.0051	0.0306 ± 0.0192
COMPAS/GBDT	0.6583 ± 0.0082	0.1688 ± 0.0177	0.6475 ± 0.0076	0.0472 ± 0.0344	0.6443 ± 0.0075	0.0131 ± 0.0085
ACS-2/LR	0.7572 ± 0.0001	0.1818 ± 0.0012	0.7544 ± 0.0005	0.0055 ± 0.0056	0.7570 ± 0.0003	0.0023 ± 0.0015
ACS-2/GBDT	0.7975 ± 0.0002	0.1195 ± 0.0005	0.7942 ± 0.0002	0.0017 ± 0.0010	0.7919 ± 0.0005	0.0010 ± 0.0008
ACS-2/MLP	0.7896 ± 0.0002	0.1192 ± 0.0007	0.7861 ± 0.0002	0.0034 ± 0.0025	0.7837 ± 0.0004	0.0012 ± 0.0008
ACS-5/LR	0.4561 ± 0.0002	0.4882 ± 0.0091	0.4526 ± 0.0004	0.0878 ± 0.0212	0.4544 ± 0.0007	0.1644 ± 0.0391
ACS-5/GBDT	0.5230 ± 0.0005	0.3991 ± 0.0072	0.5207 ± 0.0004	0.0944 ± 0.0144	0.5203 ± 0.0003	0.1401 ± 0.0121
ACS-5/MLP	0.5142 ± 0.0004	0.4854 ± 0.0069	0.5121 ± 0.0004	0.1018 ± 0.0105	0.5117 ± 0.0003	0.1898 ± 0.0456
BiasBios/BERT	0.8160 ± 0.0009	0.4167 ± 0.0166	0.8139 ± 0.0006	0.2881 ± 0.0180	0.7712 ± 0.0012	0.2650 ± 0.0145

E.2 Additional learned post-processors

We additionally compare with FRAPPÉ (Tifrea et al., 2023) and Modification Bias Score (MBS) (Chen et al., 2024) on Adult with frozen LR and GBDT predictors and the same five test splits. Our FRAPPÉ implementation learns a linear logit correction with a predictive-preservation KL term and a Gaussian-kernel MMD penalty on positive-label examples. Its regularization strength is selected on an inner 80/20 partition of the post-processing data, followed by refitting on the full post-processing split. MBS uses the binary equal-opportunity specialization of the bias score, with a one-hidden-layer, eight-unit joint (A, Y) estimator fitted on the pretraining data and a modification threshold selected on the post-processing split. Both implementations use PyTorch; selection minimizes validation EOpp subject to a two-percentage-point accuracy-loss allowance. The FLA operating points are those of the main table.

Table 6 shows that FLA attains lower mean EOpp disparity and higher mean macro F1 in both settings, while the learned post-processors retain higher accuracy. These results illustrate the different utility–fairness trade-offs available without retraining the base predictor.

Table 6: Additional post-processors on Adult: mean ± sample standard deviation over five seeds.

Model	Method	Accuracy	Macro F1	ΔEOpp
LR	Base	0.852 ± 0.002	0.782 ± 0.003	0.102 ± 0.012
LR	FLA	0.844 ± 0.002	0.791 ± 0.003	0.019 ± 0.008
LR	FRAPPÉ-linear	0.851 ± 0.002	0.781 ± 0.003	0.041 ± 0.036
LR	MBS-EOp	0.850 ± 0.002	0.779 ± 0.004	0.029 ± 0.016
GBDT	Base	0.871 ± 0.001	0.812 ± 0.002	0.105 ± 0.010
GBDT	FLA	0.858 ± 0.001	0.813 ± 0.001	0.013 ± 0.008
GBDT	FRAPPÉ-linear	0.868 ± 0.003	0.809 ± 0.008	0.036 ± 0.026
GBDT	MBS-EOp	0.868 ± 0.004	0.802 ± 0.012	0.024 ± 0.012

E.3 Ratio-based fairness diagnostics

For binary tasks, we also report the ratio of the smaller to the larger group true-positive rate (EOpp ratio) and the corresponding ratio of positive-prediction rates (DP ratio); larger values indicate closer parity. Table 7 uses the same validation-selected FLA operating points and test splits as the main results. Both ratios improve in all seven settings. On Adult, the DP ratio improves even though the absolute SP gap increases slightly: both groups' positive prediction rates increase, changing the scale against which their difference is measured. Reporting both views makes this distinction visible.

The fine-grid runs stored scalar metrics rather than groupwise confusion matrices. We reconstruct the required group rates from these records and validate the reconstruction against 873 identifiable coarse-grid records (maximum absolute rate error below 1.1×10^{-12}). None of the reported ratios has a zero denominator.

Table 7: Ratio diagnostics: mean ± sample standard deviation over five seeds, with 1 indicating parity.

	EOpp ratio		DP ratio	
Setting	Base	FLA	Base	FLA
ACS-2 / GBDT	0.845 ± 0.001	0.999 ± 0.001	0.578 ± 0.001	0.772 ± 0.004
ACS-2 / LR	0.744 ± 0.002	0.997 ± 0.002	0.516 ± 0.001	0.812 ± 0.004
ACS-2 / MLP	0.842 ± 0.001	0.999 ± 0.001	0.578 ± 0.001	0.783 ± 0.004
Adult / GBDT	0.844 ± 0.013	0.983 ± 0.010	0.291 ± 0.009	0.405 ± 0.018
Adult / LR	0.834 ± 0.019	0.973 ± 0.011	0.296 ± 0.010	0.411 ± 0.018
COMPAS / GBDT	0.746 ± 0.028	0.979 ± 0.014	0.706 ± 0.048	0.951 ± 0.045
COMPAS / LR	0.711 ± 0.027	0.958 ± 0.026	0.605 ± 0.036	0.950 ± 0.039

E.4 Fixed-strength prior correction

Table 8 compares the base predictor, fixed-strength FLA ($\tau = 1$), and the validation-selected operating points used in Table 1, on the same five evaluation splits. Fixed-strength correction lowers mean ΔEOpp in all 11 main settings, with relative reductions of 34–90% (median 64%). Accuracy decreases by less than one percentage point in ten settings; on BiasBios it decreases by 1.61 points, while macro F1 decreases from 0.8160 to 0.7782. Thus the fixed correction already improves recall parity, with the utility cost varying across tasks. On ACSIncome-2, validation selection further reduces disparity at a larger accuracy cost, whereas $\tau = 1$ retains more of the base predictor's utility. Selection on validation data need not yield a lower test disparity than $\tau = 1$ in every setting, as illustrated by COMPAS/LR and ACSIncome-5/GBDT. Results on the sparse Communities stress test are reported in Appendix J.

Table 8: Fixed and validation-selected prior correction: mean ± standard deviation over five evaluation seeds. $\hat{\tau}$ is selected by the main validation protocol; Base corresponds to $\tau = 0$. All rows use the same frozen base predictors and evaluation splits.

Setting	Adjustment	Accuracy	Macro F1	ΔEOpp
Adult/LR	Base	0.8519 ± 0.0022	0.7823 ± 0.0031	0.1020 ± 0.0121
Adult/LR	$\tau = 1$	0.8439 ± 0.0022	0.7912 ± 0.0029	0.0195 ± 0.0104
Adult/LR	$\hat{\tau} = 0.988$	0.8439 ± 0.0023	0.7908 ± 0.0030	0.0195 ± 0.0083
Adult/GBDT	Base	0.8705 ± 0.0010	0.8123 ± 0.0021	0.1046 ± 0.0095
Adult/GBDT	$\tau = 1$	0.8609 ± 0.0005	0.8144 ± 0.0017	0.0156 ± 0.0096
Adult/GBDT	$\hat{\tau} = 1.107$	0.8585 ± 0.0008	0.8130 ± 0.0012	0.0128 ± 0.0080
COMPAS/LR	Base	0.6820 ± 0.0077	0.6818 ± 0.0078	0.2242 ± 0.0230
COMPAS/LR	$\tau = 1$	0.6825 ± 0.0050	0.6822 ± 0.0051	0.0231 ± 0.0193
COMPAS/LR	$\hat{\tau} = 1.349$	0.6745 ± 0.0053	0.6742 ± 0.0051	0.0306 ± 0.0192
COMPAS/GBDT	Base	0.6593 ± 0.0084	0.6583 ± 0.0082	0.1688 ± 0.0177
COMPAS/GBDT	$\tau = 1$	0.6510 ± 0.0090	0.6504 ± 0.0087	0.0942 ± 0.0168
COMPAS/GBDT	$\hat{\tau} = 1.936$	0.6449 ± 0.0078	0.6443 ± 0.0075	0.0131 ± 0.0085
ACS-2/LR	Base	0.7799 ± 0.0002	0.7572 ± 0.0001	0.1818 ± 0.0012
ACS-2/LR	$\tau = 1$	0.7775 ± 0.0002	0.7626 ± 0.0002	0.0663 ± 0.0016
ACS-2/LR	$\hat{\tau} = 1.676$	0.7677 ± 0.0004	0.7570 ± 0.0003	0.0023 ± 0.0015
ACS-2/GBDT	Base	0.8129 ± 0.0003	0.7975 ± 0.0002	0.1195 ± 0.0005
ACS-2/GBDT	$\tau = 1$	0.8087 ± 0.0003	0.7975 ± 0.0003	0.0494 ± 0.0010
ACS-2/GBDT	$\hat{\tau} = 1.805$	0.8003 ± 0.0005	0.7919 ± 0.0005	0.0010 ± 0.0008
ACS-2/MLP	Base	0.8061 ± 0.0002	0.7896 ± 0.0002	0.1192 ± 0.0007
ACS-2/MLP	$\tau = 1$	0.8022 ± 0.0003	0.7903 ± 0.0002	0.0490 ± 0.0010
ACS-2/MLP	$\hat{\tau} = 1.845$	0.7922 ± 0.0005	0.7837 ± 0.0004	0.0012 ± 0.0008
ACS-5/LR	Base	0.4646 ± 0.0003	0.4561 ± 0.0002	0.4882 ± 0.0091
ACS-5/LR	$\tau = 1$	0.4587 ± 0.0005	0.4534 ± 0.0008	0.1752 ± 0.0441
ACS-5/LR	$\hat{\tau} = 0.921$	0.4596 ± 0.0003	0.4544 ± 0.0007	0.1644 ± 0.0391
ACS-5/GBDT	Base	0.5214 ± 0.0005	0.5230 ± 0.0005	0.3991 ± 0.0072
ACS-5/GBDT	$\tau = 1$	0.5168 ± 0.0005	0.5184 ± 0.0005	0.1335 ± 0.0350
ACS-5/GBDT	$\hat{\tau} = 0.795$	0.5185 ± 0.0004	0.5203 ± 0.0003	0.1401 ± 0.0121
ACS-5/MLP	Base	0.5123 ± 0.0004	0.5142 ± 0.0004	0.4854 ± 0.0069
ACS-5/MLP	$\tau = 1$	0.5088 ± 0.0005	0.5112 ± 0.0003	0.1979 ± 0.0607
ACS-5/MLP	$\hat{\tau} = 0.92$	0.5093 ± 0.0005	0.5117 ± 0.0003	0.1898 ± 0.0456
BiasBios/BERT	Base	0.8638 ± 0.0005	0.8160 ± 0.0009	0.4167 ± 0.0166
BiasBios/BERT	$\tau = 1$	0.8477 ± 0.0006	0.7782 ± 0.0009	0.2762 ± 0.0177
BiasBios/BERT	$\hat{\tau} = 1.075$	0.8444 ± 0.0006	0.7712 ± 0.0012	0.2650 ± 0.0145

Table 9: Companion diagnostics for Table 1: binary and ACSIncome-2 tasks.

Dataset	Model	Method	Param.	Acc	ΔSP	ΔEO
Adult	LR	Base	—	0.852	0.179	0.102
Adult	LR	Hardt	—	0.837	0.089	0.037
Adult	LR	FairProj.	t=0.8	0.851	0.154	0.060
Adult	LR	LP-EOpp	α=0.005	0.850	0.142	0.052
Adult	LR	FLA	τ=0.988	0.844	0.188	0.088
Adult	GBDT	Base	—	0.871	0.188	0.105
Adult	GBDT	Hardt	—	0.857	0.102	0.037
Adult	GBDT	FairProj.	t=1	0.870	0.154	0.048
Adult	GBDT	LP-EOpp	α=0.02	0.870	0.156	0.049
Adult	GBDT	FLA	τ=1.11	0.858	0.198	0.086
COMPAS	LR	Base	—	0.682	0.243	0.224
COMPAS	LR	Hardt	—	0.666	0.034	0.080
COMPAS	LR	FairProj.	t=0.8	0.681	0.038	0.084
COMPAS	LR	LP-EOpp	α=0.005	0.679	0.035	0.098
COMPAS	LR	FLA	τ=1.35	0.675	0.027	0.131
COMPAS	GBDT	Base	—	0.659	0.151	0.169
COMPAS	GBDT	Hardt	—	0.637	0.053	0.074
COMPAS	GBDT	FairProj.	t=1	0.645	0.028	0.127
COMPAS	GBDT	LP-EOpp	α=0.001	0.648	0.036	0.105
COMPAS	GBDT	FLA	τ=1.94	0.645	0.025	0.132
ACSIncome-2	LR	Base	—	0.780	0.204	0.182
ACSIncome-2	LR	Hardt	—	0.774	0.075	0.006
ACSIncome-2	LR	FairProj.	t=1	0.777	0.084	0.010
ACSIncome-2	LR	LP-EOpp	α=0.04	0.777	0.084	0.010
ACSIncome-2	LR	FLA	τ=1.68	0.768	0.087	0.011
ACSIncome-2	GBDT	Base	—	0.813	0.188	0.120
ACSIncome-2	GBDT	Hardt	—	0.804	0.086	0.005
ACSIncome-2	GBDT	FairProj.	t=1	0.810	0.102	0.019
ACSIncome-2	GBDT	LP-EOpp	α=0.005	0.810	0.098	0.016
ACSIncome-2	GBDT	FLA	τ=1.8	0.800	0.110	0.029
ACSIncome-2	MLP	Base	—	0.806	0.185	0.119
ACSIncome-2	MLP	Hardt	—	0.797	0.082	0.005
ACSIncome-2	MLP	FairProj.	t=1	0.803	0.100	0.021
ACSIncome-2	MLP	LP-EOpp	α=0.001	0.803	0.100	0.021
ACSIncome-2	MLP	FLA	τ=1.84	0.792	0.105	0.024

F MAIN-TABLE COMPANION METRICS

Tables 9 and 10 report accuracy, ΔSP, and ΔEO for the same selected operating points as Table 1. This keeps the main table focused on macro F1-Score and the primary EOpp metric while preserving the utility and cross-metric fairness diagnostics.

G METRIC-SPECIFIC SELECTION DIAGNOSTICS

Tables 11 and 12 repeat the same accuracy-constrained validation rule as the main protocol while selecting τ by ΔEO and ΔSP, respectively. These tables are diagnostic: they show how the same closed-form FLA sweep can be selected for different fairness criteria, while LP must solve the corresponding constrained problem separately.

Table 10: Companion diagnostics for Table 1: ACSIncome-5 and BiasBios.

Dataset	Model	Method	Param.	Acc	ΔSP	ΔEO
ACSIncome-5	LR	Base	—	0.465	0.371	0.488
ACSIncome-5	LR	Hardt	—	—	—	—
ACSIncome-5	LR	FairProj.	t=0.6	0.464	0.217	0.208
ACSIncome-5	LR	LP-EOpp	α=0.02	0.462	0.135	0.092
ACSIncome-5	LR	FLA	τ=0.921	0.460	0.147	0.166
ACSIncome-5	GBDT	Base	—	0.521	0.293	0.399
ACSIncome-5	GBDT	Hardt	—	—	—	—
ACSIncome-5	GBDT	FairProj.	t=0.8	0.520	0.172	0.144
ACSIncome-5	GBDT	LP-EOpp	α=0.04	0.519	0.122	0.125
ACSIncome-5	GBDT	FLA	τ=0.795	0.518	0.165	0.143
ACSIncome-5	MLP	Base	—	0.512	0.307	0.485
ACSIncome-5	MLP	Hardt	—	—	—	—
ACSIncome-5	MLP	FairProj.	t=0.6	0.512	0.194	0.223
ACSIncome-5	MLP	LP-EOpp	α=0.06	0.510	0.130	0.103
ACSIncome-5	MLP	FLA	τ=0.92	0.509	0.146	0.190
BiasBios	BERT	Base	—	0.864	0.083	0.417
BiasBios	BERT	Hardt	—	—	—	—
BiasBios	BERT	FairProj.	t=1	0.858	0.082	0.196
BiasBios	BERT	LP-EOpp	α=0.15	0.863	0.083	0.288
BiasBios	BERT	FLA	τ=1.07	0.844	0.075	0.265

Table 11: Cross-metric diagnostics for EO-selected FLA and LP-EO.

Dataset	Model	Method	Param.	Acc	Macro F1-Score	ΔSP	ΔEOpp	ΔEO
Adult	LR	FLA	τ=0.454	0.850	0.789	0.183	0.049	0.081
Adult	LR	LP-EO	—	0.838	—	0.093	0.028	0.028
Adult	GBDT	FLA	τ=0.59	0.867	0.817	0.191	0.035	0.077
Adult	GBDT	LP-EO	—	0.858	—	0.104	0.024	0.024
COMPAS	LR	FLA	τ=1.06	0.682	0.682	0.023	0.023	0.082
COMPAS	LR	LP-EO	—	0.683	—	0.080	0.068	0.085
COMPAS	GBDT	FLA	τ=1.29	0.649	0.648	0.043	0.068	0.088
COMPAS	GBDT	LP-EO	—	0.631	—	0.047	0.037	0.063
ACSIncome-2	LR	FLA	τ=1.73	0.767	0.756	0.082	0.008	0.009
ACSIncome-2	LR	LP-EO	—	0.778	—	0.092	0.016	0.017
ACSIncome-2	GBDT	FLA	τ=2	0.798	0.790	0.100	0.011	0.019
ACSIncome-2	GBDT	LP-EO	—	0.805	—	0.084	0.006	0.006
ACSIncome-2	MLP	FLA	τ=2	0.790	0.782	0.097	0.009	0.017
ACSIncome-2	MLP	LP-EO	—	0.797	—	0.085	0.002	0.003
ACSIncome-5	LR	FLA	τ=0.889	0.460	0.455	0.155	0.160	0.160
ACSIncome-5	LR	LP-EO	—	0.461	—	0.125	0.077	0.085
ACSIncome-5	GBDT	FLA	τ=0.788	0.518	0.520	0.166	0.142	0.144
ACSIncome-5	GBDT	LP-EO	α=0.06	0.520	—	0.132	0.095	0.105
ACSIncome-5	MLP	FLA	τ=0.92	0.509	0.512	0.146	0.190	0.190
ACSIncome-5	MLP	LP-EO	—	—	—	—	—	—
BiasBios	BERT	FLA	τ=1.07	0.844	0.771	0.075	0.265	0.265
BiasBios	BERT	LP-EO	α=0.1	0.862	—	0.083	0.242	0.242

Table 12: Cross-metric diagnostics for SP-selected FLA and LP-SP.

Dataset	Model	Method	Param.	Acc	Macro F1-Score	ΔSP	ΔEOpp	ΔEO
Adult	LR	FLA	τ=0.015	0.852	0.783	0.178	0.095	0.095
Adult	LR	LP-SP	α=0.001	0.834	—	0.013	0.349	0.349
Adult	GBDT	FLA	τ=0	0.871	0.812	0.188	0.105	0.105
Adult	GBDT	LP-SP	α=0.001	0.851	—	0.010	0.341	0.341
COMPAS	LR	FLA	τ=1.37	0.674	0.674	0.031	0.036	0.135
COMPAS	LR	LP-SP	α=0.02	0.679	—	0.037	0.038	0.121
COMPAS	GBDT	FLA	τ=1.83	0.645	0.645	0.021	0.021	0.123
COMPAS	GBDT	LP-SP	α=0.001	0.646	—	0.032	0.034	0.113
ACSIncome-2	LR	FLA	τ=2	0.761	0.752	0.060	0.031	0.031
ACSIncome-2	LR	LP-SP	α=0.001	0.769	—	0.004	0.108	0.108
ACSIncome-2	GBDT	FLA	τ=2	0.798	0.790	0.100	0.011	0.019
ACSIncome-2	GBDT	LP-SP	α=0.001	0.801	—	0.004	0.130	0.130
ACSIncome-2	MLP	FLA	τ=2	0.790	0.782	0.097	0.009	0.017
ACSIncome-2	MLP	LP-SP	α=0.001	0.794	—	0.003	0.128	0.128
ACSIncome-5	LR	FLA	τ=1.14	0.457	0.451	0.122	0.202	0.202
ACSIncome-5	LR	LP-SP	α=0.001	0.456	—	0.028	0.147	0.163
ACSIncome-5	GBDT	FLA	τ=1.32	0.514	0.515	0.107	0.204	0.253
ACSIncome-5	GBDT	LP-SP	α=0.005	0.512	—	0.030	0.166	0.185
ACSIncome-5	MLP	FLA	τ=1.26	0.507	0.509	0.123	0.237	0.241
ACSIncome-5	MLP	LP-SP	α=0.001	0.503	—	0.031	0.177	0.194
BiasBios	BERT	FLA	τ=1.08	0.844	0.771	0.075	0.265	0.265
BiasBios	BERT	LP-SP	α=0.015	0.846	—	0.018	0.274	0.274

H IMPROVEMENT-PATTERN CORRELATION: FLA VS. LP-EOPP

Table 13: Per dataset–model pair: base-model ΔEOpp and improvement (Base−Method) for FLA (val-selected τ) and LP-EOpp (val-selected α). Spearman rank correlation between FLA and LP improvement vectors: $\rho = 0.99$, $p<10^{-4}$, $n = 11$. The correlation describes similarity in the improvement patterns. Communities is reported separately: FLA improvement is zero (val-selection chooses $\tau\approx0$), while LP-EOpp improves on the base (see Appendix J).

Dataset	Model	Base ΔEOpp	FLA improv.	LP-EOpp improv.
Adult	LR	0.102	0.083	0.084
Adult	GBDT	0.105	0.092	0.088
COMPAS	LR	0.224	0.194	0.195
COMPAS	GBDT	0.169	0.156	0.122
ACSIncome-2	LR	0.182	0.180	0.176
ACSIncome-2	GBDT	0.120	0.119	0.118
ACSIncome-2	MLP	0.119	0.118	0.116
ACSIncome-5	LR	0.488	0.324	0.400
ACSIncome-5	GBDT	0.399	0.259	0.305
ACSIncome-5	MLP	0.485	0.296	0.384
BiasBios	BERT	0.417	0.152	0.129
			Spearman $\rho = 0.99, p = 0.0000, n = 11$	

I IMPLEMENTATION DETAILS AND HYPERPARAMETERS

Tabular models. Logistic regression: `sklearn LogisticRegression`, `max_iter=10000`. GBDT: `LightGBM LGBMClassifier`, 100 estimators. MLP: 3-layer ReLU network (256–128–K), Adam, lr=1e-3, batch=256, 50 epochs.

BiasBios BERT. `bert-base-uncased` fine-tuned for 3 epochs, Adam lr=2e-5, batch=32, gradient accumulation=4, linear warmup (10% steps).

FLA post-hoc. $\pi_{y|a}$ estimated from the post-processing validation split (no test leakage). τ swept over 2001 values in $[0, 2]$; selected on the post-processing validation set (minimizing ΔEOpp at matched accuracy); wall-clock fit time < 1 s for all tabular settings and < 5 ms for BiasBios.

Hardware. Tabular experiments: Intel Xeon CPU. BiasBios BERT fine-tuning: single NVIDIA GPU. LP uses the default solver selected by `cvxpy`.

J COMMUNITIES: PRIOR-ESTIMATION FAILURE AND FLA'S SCOPE

Communities & Crime is a 5-class task (K=5, G=2, $N\approx$2,000). Several group–class cells contain fewer than 20 samples, limiting the data available to estimate their conditional priors.

Table 14: Communities: all post-hoc methods for LR and GBDT predictors. Best ΔEOpp per predictor in **bold**.

Method	Logistic Reg.		GBDT	
	Acc	ΔEOpp↓	Acc	ΔEOpp↓
Base	0.671	0.546	0.678	0.753
FLA (val-sel. $\tau\approx$0)	0.671	0.546	0.678	0.753
FLA (τ=1)	0.466	0.941	0.483	1.000
LP-EOpp (best α)	0.667	**0.492**	0.672	**0.414**
Hardt-EOpp	0.707	1.000	0.697	1.000
FairProjection-EOpp	0.628	0.633	0.619	0.733

Results in the sparse-cell setting. At $\tau = 1$, FLA increases disparity for both predictors. Validation instead selects $\tau \approx 0$, retaining the base classifier's performance. LP-EOpp reduces ΔEOpp to 0.492 (LR) and 0.414 (GBDT). The limited cell support makes this a useful stress case for an approach that estimates group-conditional priors. Appendix K examines sensitivity to fit-set size using the existing subsampling results.

K PRIOR-ESTIMATION ROBUSTNESS

FLA's adjustment relies on the empirical prior $\hat{\pi}_{y|a}$ estimated from the post-processing fit set. We ablate robustness by subsampling the fit set to fractions $r \in \{100\%, 50\%, 25\%, 10\%, 5\%\}$ and measuring ΔEOpp at $\tau = 1$ across five seeds.

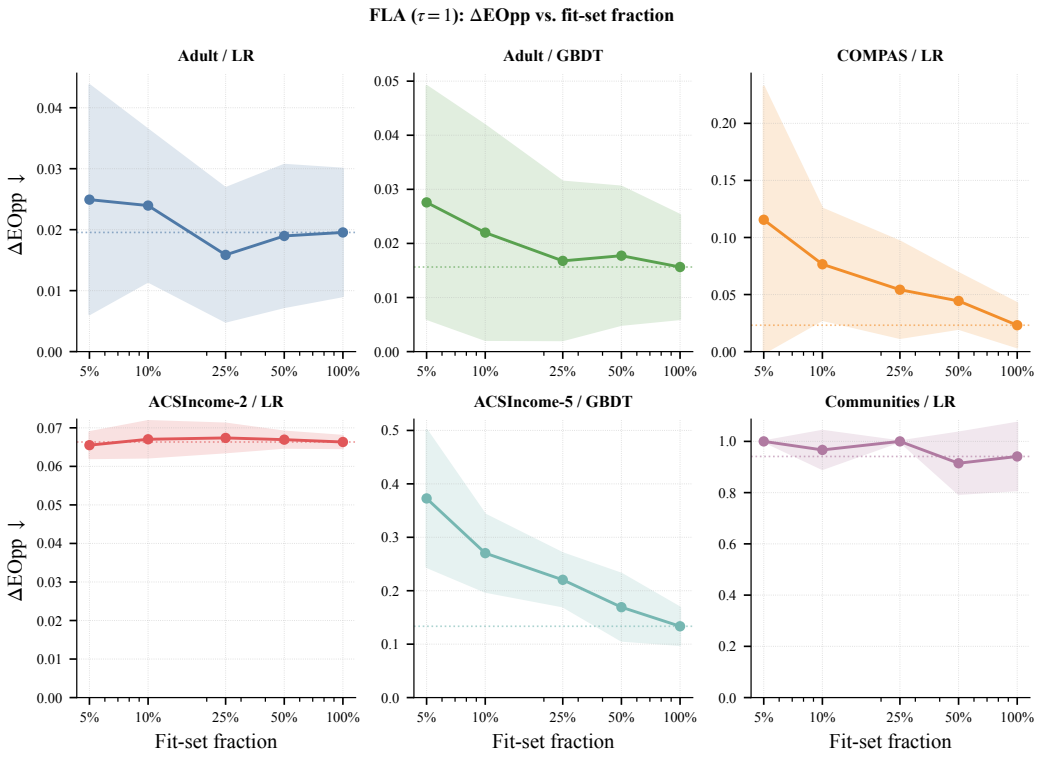

Figure 5: ΔEOpp (τ=1) vs. fit-set fraction across datasets. Shaded bands show ±1 std across 5 seeds. Adult and ACSIncome-2 change little over the evaluated fit-set fractions. ACSIncome-5 shows greater sensitivity to subsampling. Communities remains near ΔEOpp≈ 1.0 across fractions, with a minimum cell count of three even in the full fit set.

Key findings. (1) Binary datasets with KG=4 cells (Adult, COMPAS) show $\leq 5\times$ ΔEOpp increase from 100% to 5% fit set, with absolute ΔEOpp remaining small (< 0.12). (2) ACSIncome-2 (N>50k) is essentially unaffected across all fractions (ΔEOpp changes by <2%). (3) ACSIncome-5 (KG=25 cells) shows 2.8× degradation (0.133 → 0.373) as the fit set is reduced. (4) Communities stays flat near ΔEOpp=1.0 across fractions: the full split already has sparse cells, and reducing the fit set does not improve its performance.

Connection to per-cell sample size. The observed sensitivity to fit-set size is consistent with the role of cell support in prior estimation. For n_a independent labels from group a, the binomial variance of $\hat{\pi}_{y|a}$ is $\pi_{y|a}(1-\pi_{y|a})/n_a$. When the cell is sufficiently well sampled for a delta-method approximation, $\mathrm{Var}(\log \hat{\pi}_{y|a}) \approx (1-\pi_{y|a})/(n_a \pi_{y|a})$. This explains why small expected cell counts can make log-prior estimates unstable. The subsampling results illustrate empirical sensitivity to the data available for estimating these priors.

L Broader Impact

Positive impact. FLA lowers the barrier to deploying fair classifiers in high-stakes domains such as hiring, credit scoring, and clinical triage. Because it requires no model retraining and no solver infrastructure, it is accessible to practitioners who cannot modify production models or afford LP-based post-processing pipelines. The adjustment uses estimated group-conditional priors ($\hat{\pi}_{y|a}$) and a single hyperparameter τ selected on a validation split.

Potential negative impact. Post-hoc FLA uses the sensitive attribute a at inference time, so deployment requires that this attribute be available and its use appropriate to the application. Over-relying on ΔEOpp/GBER as the sole fairness criterion may obscure other disparities (individual

fairness, intersectional groups); practitioners should complement FLA with broader audits. Finally, improving group-level recall parity does not address upstream biases in training labels or feature representations.

Data and privacy. All datasets used in this work (Adult, COMPAS, Communities, ACSIncome, BiasBios) are publicly available benchmarks with established use in the fairness literature. Sensitive attributes (sex, race) are used for prior estimation, group-aware prediction, and fairness evaluation. Practitioners deploying FLA should ensure that attribute collection and retention comply with applicable data-protection regulations.

M REPRODUCIBILITY DETAILS

All tabular experiments (Adult, COMPAS, ACS Income-2, ACS Income-5, Communities) are run with 5 random seeds (33–37) using publicly available datasets. ACS Income data is obtained via the `folktables` API; COMPAS and Adult are from UCI/ProPublica; Communities from UCI. BiasBios experiments use the public dataset of De-Arteaga et al. (2019) with pretrained BERT fine-tuned for 3 epochs per seed (seeds 33–37).

Code and data. The supplementary package provides a self-contained post-hoc FLA implementation and a runnable demo. FLA is a closed-form logit-offset rule: given logits, group labels, and empirical group-class priors, the core method is reproducible from the paper's equations and the demo in roughly ten lines of NumPy. All datasets are publicly available (UCI, ProPublica, Folktables API, BiasBios); all baselines rely on standard libraries (fairlearn, cvxpy, lightgbm, scikit-learn) as documented in Appendix I. The in-processing BiasBios diagnostic requires BERT fine-tuning: one GPU, ≈ 3 hr/seed. The complete experiment codebase, result summaries, and pre-computed logits will be made publicly available on GitHub following code organization.

Software. Python 3.10+; numpy $<$2.0 (required for folktables compatibility); scikit-learn, scipy, cvxpy, fairlearn, lightgbm, torch, transformers, datasets. The supplementary `README.md` documents the minimal FLA code and public data/model sources.