GLOBAL FAIRNESS: LEARNING FAIR SCORE FUNCTIONS ACROSS ACCEPTANCE RATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Fair learning methods typically optimize the accuracy–fairness trade-off of a binary classifier at a particular decision threshold. In many applications, however, the deployment threshold is determined by a desired acceptance rate, which may be unknown during training and vary with downstream requirements. This motivates learning a score function that remains both fair and predictive across a range of acceptance rates. Existing score-level fairness measures move beyond a single threshold, but do not explicitly account for the range of acceptance rates over which group disparity persists. We introduce *global fairness*, a score-function fairness measure that aggregates threshold-wise group disparity according to the population score density. Our key observation is that the location of disparity relative to this density determines its acceptance-rate coverage: the same disparity in a high-density score region spans a wider range of acceptance rates than in a low-density region. We show that this density-weighted formulation is equivalent to averaging group disparity uniformly over population acceptance rates. Consequently, global fairness can distinguish score functions whose threshold-wise disparities have the same magnitude and shape but substantially different acceptance-rate coverage, including cases not distinguished by Kolmogorov–Smirnov or Wasserstein-1 measures. To jointly account for fairness and predictive utility, we introduce *global accuracy*, which aggregates predictive performance across acceptance rates, and formulate learning as maximizing global accuracy subject to a global fairness constraint. We establish consistency of the empirical global fairness estimator and characterize global accuracy over the full acceptance-rate range through its relationship with AUC. Experiments on three fairness benchmarks demonstrate favorable global accuracy–fairness trade-offs and more stable fairness across acceptance rates, while remaining competitive under conventional single-threshold evaluation.

1 INTRODUCTION

Fair machine learning aims to mitigate undesirable disparities in automated decision systems while preserving predictive utility. In many classification problems, a model first produces a real-valued score and then converts it into a binary decision by comparing the score with a threshold (Hardt et al., 2016; Zafar et al., 2017; Agarwal et al., 2018). Most existing fair learning methods accordingly optimize the accuracy–fairness trade-off of the binary classifier induced by a particular threshold (Hardt et al., 2016; Agarwal et al., 2018; Diana et al., 2021; Buyl et al., 2024).

In many applications, however, the deployment threshold is not known during training and may vary with downstream requirements on the acceptance rate, i.e., the proportion of individuals predicted as positive. Specifically, for a score function s and threshold t, the acceptance rate is $\alpha(t) = \mathbb{P}(s(X) \geq t)$. For example, medical decision systems may adjust diagnostic thresholds according to desired positive test rates (Pepe, 2003), while universities and hiring systems may determine enrollment or admission rates based on recruitment needs and capacity constraints (Espenshade et al., 2004; Barocas et al., 2019; Raghavan et al., 2020). Consequently, the same learned score function may induce different binary classifiers under different acceptance-rate requirements. When the deployment acceptance rate is unknown during training, ensuring fairness only at a particular threshold is therefore insufficient. Instead, we seek a *fair score function* that maintains fairness across a broad range of possible acceptance rates.

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.

To learn such a fair score function, we consider existing score-level fairness measures, which aggregate threshold-wise group disparity in different ways. The Kolmogorov–Smirnov (KS) distance captures the maximum disparity over all thresholds (Vogel et al., 2021), whereas the Wasserstein-1 distance aggregates disparity uniformly over the threshold range (Jiang et al., 2020). Related work further develops threshold-invariant procedures or score transformations to improve fairness measured by the Wasserstein-1 distance (Chen & Wu, 2020; Wei et al., 2020; Kwegyir-Aggrey et al., 2023). While these approaches extend fairness beyond a predetermined threshold, they do not account for the width of the acceptance-rate range over which group disparity persists. We illustrate this limitation with two score functions whose threshold-wise disparity curves have identical shapes and magnitudes but occur at different regions of the population score distribution. Although the disparity extends over threshold intervals of the same width, these intervals can correspond to acceptance-rate ranges of substantially different widths, meaning that the disparity persists over different portions of the possible deployment range and thus has different implications for deployment flexibility. Nevertheless, KS and Wasserstein-1 assign the same fairness value to both score functions, failing to capture this deployment-relevant distinction.

Our key observation is that the acceptance-rate coverage of a threshold interval depends on the population score density within that interval: the greater the probability mass, the wider the corresponding range of acceptance rates. Thus, for equally wide threshold intervals over which disparity occurs, an interval in a high-density region corresponds to a wider range of acceptance rates than one in a low-density region. Motivated by this observation, we propose *global fairness*, a score-function fairness measure that accounts for both the magnitude of threshold-wise disparity and the population score density where the disparity occurs. Specifically, global fairness weights threshold-wise disparity by the population score density. This weighting has a direct operational interpretation: the population score density provides exactly the change-of-variable factor between score thresholds and acceptance rates, making global fairness equivalent to the average threshold-wise disparity taken uniformly over population acceptance rates. Consequently, global fairness distinguishes score functions with identical threshold-wise disparity patterns but different acceptance-rate coverage, even when KS and Wasserstein-1 do not.

The same acceptance-rate uncertainty also motivates evaluating predictive utility across possible deployment rates. We therefore introduce *global accuracy*, which aggregates predictive performance across population acceptance rates, and formulate learning as maximizing global accuracy subject to a global fairness constraint. We develop a differentiable empirical estimator of global fairness and establish its statistical consistency. We further show that, over the full range of acceptance rates, global accuracy is an affine function of AUC, providing a direct connection between our acceptance-rate-based formulation and a standard measure of ranking performance.

Our main contributions are summarized as follows.

- We introduce *global fairness*, a density-weighted measure of threshold-wise disparity that equivalently averages disparity over population acceptance rates. This formulation accounts for both the magnitude of disparity and the acceptance-rate range over which it persists, distinguishing patterns that KS and Wasserstein-1 treat identically despite substantially different acceptance-rate coverage.
- We develop a tractable framework for learning score functions under global fairness constraints. We introduce *global accuracy*, establish consistency of the empirical global fairness estimator, and show that global accuracy over the full acceptance-rate range admits an affine characterization in terms of AUC.
- Experiments on three fairness benchmarks demonstrate favorable global accuracy–fairness trade-offs, particularly under stringent global fairness constraints, and more stable fairness across acceptance rates, while remaining competitive under conventional single-threshold evaluation.

2 Fair Score Function Learning

In fair classification, we observe samples $D := \{(X_i, Y_i, Z_i)\}_{i=1}^n$ drawn i.i.d. from an unknown distribution P over $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, where $X \in \mathcal{X} \subseteq \mathbb{R}^d$ denotes the input feature, $Y \in \mathcal{Y} := \{-1, 1\}$ is the class label, and $Z \in \mathcal{Z} := \{-1, 1\}$ is the sensitive attribute (e.g., gender or race). A standard score-based approach first learns a score function $s : \mathcal{X} \to \mathbb{R}$ and then converts it into a binary classifier by thresholding: $\widehat{Y} := h_t(X) := 2\mathbf{1}\{s(X) > t\} - 1$, where $t \in \mathbb{R}$ is a decision threshold and $\mathbf{1}\{\cdot\}$ denotes the indicator function.

Utility and Fairness at a Fixed Threshold. The predictive utility of the thresholded classifier h_t is measured by its classification accuracy

$$\mathrm{Acc}(h_t) := \mathbb{P}(h_t(X) = Y),$$

or equivalently by its classification error $\mathrm{Err}(h_t) := 1 - \mathrm{Acc}(h_t)$. Group fairness instead evaluates whether relevant decision statistics differ across sensitive groups. One widely used notion is *demographic parity* (DP) (Dwork et al., 2012; Zemel et al., 2013), which measures the difference in positive prediction rates across groups:

$$\Delta_{\mathrm{DP}}(h_t) := \left|\mathbb{P}(h_t(X) = 1 \mid Z = 1) - \mathbb{P}(h_t(X) = 1 \mid Z = -1)\right|. \tag{1}$$

Another common notion is *equality of opportunity* (EO) (Hardt et al., 2016), which measures the difference in true positive rates:

$$\Delta_{\mathrm{EO}}(h_t) := \left|\mathbb{P}(h_t(X) = 1 \mid Y = 1, Z = 1) - \mathbb{P}(h_t(X) = 1 \mid Y = 1, Z = -1)\right|. \tag{2}$$

Both notions evaluate the fairness of the binary classifier induced by a particular threshold. Changing the threshold can therefore change the resulting unfairness. When the deployment threshold is unknown during training or may vary after deployment, fairness at a single threshold does not characterize the fairness of the underlying score function across possible deployment thresholds. This motivates evaluating the score function beyond a single predetermined threshold.

Fairness Measures for Score Functions. To evaluate fairness across thresholds, we consider the group-conditional distributions of the learned scores. Let

$$F_1(t) := \mathbb{P}(s(X) \le t \mid Z = 1), \qquad F_{-1}(t) := \mathbb{P}(s(X) \le t \mid Z = -1), \qquad F(t) := \mathbb{P}(s(X) \le t).$$

At threshold t, the corresponding DP disparity in Eq. (1) can be written as

$$\begin{aligned}\Delta_{\mathrm{DP}}(h_t) &= \left|\mathbb{P}(s(X) > t \mid Z = 1) - \mathbb{P}(s(X) > t \mid Z = -1)\right| \\ &= \left|(1 - F_1(t)) - (1 - F_{-1}(t))\right| = \left|F_1(t) - F_{-1}(t)\right| =: D_s(t),\end{aligned} \tag{3}$$

where $D_s(t)$ denotes the *threshold-wise disparity function* induced by s. Evaluating the fairness of a score function requires considering $D_s(t)$ across possible thresholds rather than only at one predetermined threshold.

Existing score-level fairness measures aggregate this threshold-wise disparity in different ways. Jiang et al. (2020) employs the Wasserstein-1 distance, which in the one-dimensional score setting admits the representation

$$W_1(s) := \int_{\mathcal{T}} \Delta_{\mathrm{DP}}(h_t)\, dt = \int_{\mathcal{T}} D_s(t)\, dt, \tag{4}$$

where $\mathcal{T}$ denotes the support of $s(X)$. Wasserstein-1 therefore aggregates unfairness uniformly over the numerical score-threshold space. In contrast, Vogel et al. (2021) considers the Kolmogorov–Smirnov (KS) distance,

$$\mathrm{KS}(s) := \sup_{t \in \mathcal{T}} \Delta_{\mathrm{DP}}(h_t) = \sup_{t \in \mathcal{T}} D_s(t), \tag{5}$$

which captures the largest threshold-wise disparity. Thus, KS focuses on the worst unfairness across thresholds, whereas Wasserstein-1 aggregates unfairness uniformly over raw score thresholds.

While both metrics evaluate fairness beyond a single threshold, neither accounts for the population score distribution. In particular, KS and Wasserstein-1 are insensitive to where the same disparity pattern occurs in the score space. As shown in Section 3, however, unfairness in a high-density region can persist over a broader range of population acceptance rates than the same unfairness in a low-density region. This motivates accounting for not only threshold-wise disparity, but also how broadly the resulting unfairness persists across acceptance rates.

3 Global Fairness under Acceptance-Rate Uncertainty

When the deployment acceptance rate is unknown during training or may vary after deployment, fairness should be assessed across the range of possible acceptance rates rather than at a single predetermined threshold. In this section, we show that the population score distribution determines how threshold-wise unfairness translates into its coverage across acceptance rates, which in turn affects the flexibility available for fair deployment. This distinction is not captured by existing score-level measures such as KS and Wasserstein-1. We therefore introduce *global fairness* to directly account for it.

Acceptance-Rate Ranges Depend on Score Density. Let $F(\cdot)$ and $f(\cdot)$ denote the CDF and density of the population score $s(X)$, respectively. A threshold t induces the population acceptance rate

$$\alpha(t) = \mathbb{P}(s(X) > t) = 1 - F(t).$$

Suppose that substantial unfairness occurs over a threshold interval $[t_1, t_2]$, where $t_1 < t_2$. The corresponding range of acceptance rates affected by this unfairness has width

$$\underbrace{\alpha(t_1) - \alpha(t_2)}_{\text{Width of unfair acceptance-rate range}} = F(t_2) - F(t_1) = \int_{t_1}^{t_2} \underbrace{f(t)}_{\text{Population score density}} dt. \tag{6}$$

We refer to $[\alpha(t_2), \alpha(t_1)]$ as the *unfair acceptance-rate range*: choosing a deployment acceptance rate within this range corresponds to a threshold at which substantial unfairness occurs.

Eq. (6) reveals the role of the population score distribution. The width of the unfair acceptance-rate range is exactly the population mass within the corresponding threshold interval. Thus, the same threshold-wise unfairness occurring in a high-density score region can extend over a much broader range of acceptance rates than when it occurs in a low-density region. This difference has a direct deployment implication: when the final acceptance rate is not known in advance, broader coverage of unfairness leaves less flexibility in choosing an acceptance rate that avoids substantial group disparity.

This situation arises naturally when the score function is learned before the final acceptance rate is determined. For example, a university may rank applicants using a learned score while determining its final acceptance rate according to enrollment capacity. If unfairness persists over a broad range of acceptance rates, unfair admission outcomes may arise under many possible enrollment levels. In contrast, unfairness confined to a narrow range leaves greater flexibility in choosing an enrollment level with little group disparity.

A Motivating Example. We illustrate this distinction using two stylized score functions with identical threshold-wise disparity patterns located in regions of substantially different population score density.

Example 3.1. *Let $Z \in \{-1, 1\}$ with $\mathbb{P}(Z = 1) = \mathbb{P}(Z = -1) = 1/2$. For $c \in \{30, 70\}$, consider the score function s_c with*

$$s_c(X) \mid Z = 1 \sim 0.1\,\mathcal{N}(c - 10, 9) + 0.8\,\mathcal{N}(30, 9) + 0.1\,\mathcal{N}(90, 9),$$
$$s_c(X) \mid Z = -1 \sim 0.1\,\mathcal{N}(c + 10, 9) + 0.8\,\mathcal{N}(30, 9) + 0.1\,\mathcal{N}(90, 9).$$

The corresponding threshold-wise DP disparity functions are $D_c(t) := D_{s_c}(t) = 0.1\left[\Phi\left(\frac{t-(c-10)}{3}\right) - \Phi\left(\frac{t-(c+10)}{3}\right)\right]$, where $\Phi(\cdot)$ is the CDF of Gaussian distribution. Hence, D_{30} and D_{70} have exactly the same magnitude and shape and differ only by horizontal translation:

$$D_{70}(t) = D_{30}(t - 40).$$

Despite their identical disparity patterns up to translation, s_{30} and s_{70} differ substantially in how broadly their unfairness persists across population acceptance rates. As illustrated in Figure 1, D_{30} is mainly concentrated in a high-density score region, whereas D_{70} lies primarily in a low-density region. Consequently, the unfairness associated with s_{30} extends over a substantially broader range of acceptance rates than that associated with s_{70}. Hence, s_{30} leaves less flexibility in choosing a deployment acceptance rate that avoids substantial group disparity. In the university-admission example, s_{30} may therefore yield unfair admission outcomes under a much broader range of possible enrollment levels, despite having the same threshold-wise disparity pattern as s_{70} up to translation.

Limitations of KS and Wasserstein Distances. The deployment-relevant difference above is invisible to both KS and Wasserstein-1. Since translation preserves the maximum of the disparity function,

$$\mathrm{KS}(s_{30}) = \mathrm{KS}(s_{70}).$$

Similarly, translation preserves the area under the disparity function, yielding

$$W_1(s_{30}) = W_1(s_{70}).$$

Thus, KS captures the maximum threshold-wise disparity, while Wasserstein-1 aggregates disparity uniformly over raw score thresholds. Neither accounts for how the population score distribution translates threshold-wise unfairness into its coverage across acceptance rates and, consequently, the flexibility available for fair deployment.

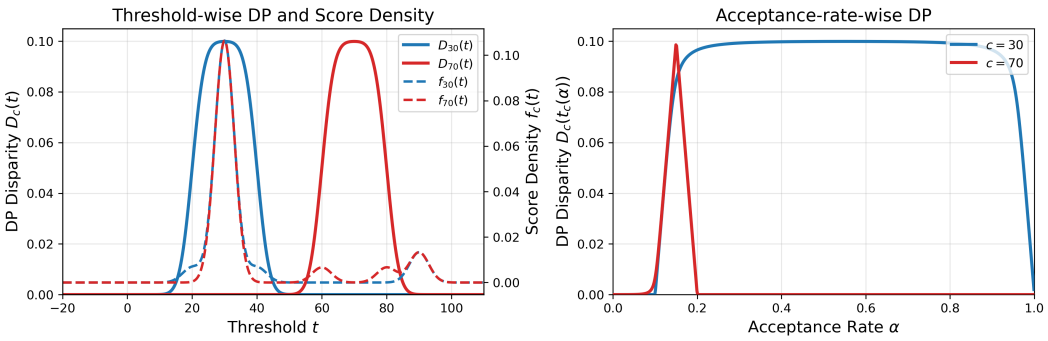


Figure 1: **The same threshold-wise disparity can correspond to substantially different unfair acceptance-rate ranges. Left:** s_{30} and s_{70} induce threshold-wise DP disparity functions with identical magnitude and shape that differ only by a translation of 40 score units. D_{30} is mainly concentrated over $[10, 50]$, where it substantially overlaps with a high-density region of the population score distribution, whereas D_{70} is mainly concentrated over $[50, 90]$, where the population score density is substantially lower. **Right:** When indexed by population acceptance rate, the unfairness associated with s_{30} extends approximately over $\alpha \in [0.1, 1]$, whereas that associated with s_{70} is concentrated approximately over $\alpha \in [0.1, 0.2]$. Thus, identical threshold-wise disparity patterns can affect substantially different ranges of deployment acceptance rates and therefore leave different degrees of flexibility for fair deployment.

Global Fairness. Motivated by this limitation, we seek a score-function fairness measure that accounts for both the magnitude of threshold-wise unfairness and how broadly it persists across population acceptance rates. We define *global fairness* by weighting threshold-wise disparity according to the population score density.

Definition 3.2 (Global Fairness)**.** *Let $D_s(t) := |F_1(t) - F_{-1}(t)|$ denote the threshold-wise DP disparity induced by a score function s. We define its global fairness gap as*

$$\Delta_G(s) := \int_{\mathcal{T}} D_s(t) f(t)\, dt, \tag{7}$$

where f denotes the density of the population score distribution and $\mathcal{T}$ its support.

The population-density weighting directly accounts for acceptance-rate coverage: threshold-wise disparity in a region containing more population mass corresponds to unfairness over a broader range of acceptance rates. More fundamentally, this weighting admits an exact acceptance-rate interpretation.

Theorem 3.3 (Acceptance-rate representation of global fairness)**.** *Let $\Delta_G(s)$ be defined as in Definition 3.2, and let $\alpha(t) = 1 - F(t)$ denote the population acceptance rate at threshold t. Then*

$$\Delta_G(s) = \int_0^1 D_s\big(t(\alpha)\big)\, d\alpha,$$

where $t(\alpha) := F^{-1}(1-\alpha)$ denotes the threshold corresponding to acceptance rate α. Thus, global fairness is equivalently the average threshold-wise disparity over population acceptance rates.

Theorem 3.3 reveals the key distinction from existing score-level fairness measures. KS considers the worst threshold, and Wasserstein-1 aggregates disparity uniformly over raw score thresholds; global fairness instead aggregates disparity uniformly over population acceptance rates. The density weighting in Eq. (7) is therefore not an arbitrary preference for high-density score regions, but precisely the change-of-variable factor from score thresholds to acceptance rates. Population density determines how threshold-wise unfairness translates into its coverage across acceptance rates, which in turn affects the flexibility available for choosing a deployment acceptance rate with little group disparity. Global fairness directly captures this deployment-relevant distinction by accounting for both the magnitude of unfairness and how broadly it persists across possible acceptance rates.

Returning to Example 3.1, global fairness distinguishes the two score functions in accordance with their acceptance-rate behavior:

$$\Delta_G(s_{30}) > \Delta_G(s_{70}),$$

Table 1: Comparison of score-level fairness measures and their behavior on the motivating example.

Metric	Aggregation of $D_s(t)$	What it captures	s_{30} vs. s_{70}
KS	$\sup_t D_s(t)$	Maximum disparity	Equal
Wasserstein-1	$\int D_s(t)\,dt$	Disparity over raw thresholds	Equal
Global Fairness	$\int D_s(t)f(t)\,dt$	Disparity over acceptance rates	$\boldsymbol{s_{30} > s_{70}}$

because the unfairness associated with s_{30} persists over a substantially broader range of population acceptance rates. In contrast, $\mathrm{KS}(s_{30}) = \mathrm{KS}(s_{70})$ and $W_1(s_{30}) = W_1(s_{70})$.

Table 1 summarizes the distinction. KS measures the largest threshold-wise disparity, Wasserstein-1 aggregates disparity uniformly over score thresholds, and global fairness aggregates disparity uniformly over population acceptance rates. By capturing the acceptance-rate coverage of unfairness, global fairness reflects how much flexibility remains for fair deployment when the final acceptance rate is uncertain.

4 Learning Algorithm

Having established global fairness as a deployment-relevant measure under acceptance-rate uncertainty, we now turn to learning a score function that is both globally fair and predictive. We first introduce *global accuracy* to quantify predictive utility across acceptance rates, complementing global fairness. We then derive threshold-independent formulations of both quantities, yielding a tractable learning objective without explicitly enumerating deployment thresholds.

Global Accuracy and Learning Objective. For an acceptance rate α, let $t(\alpha)$ denote the corresponding score threshold and define the induced classifier $h_{t(\alpha)}(x) := 2\mathbf{1}\{s(x) \geq t(\alpha)\} - 1$. Recall that global fairness admits the acceptance-rate representation

$$\Delta_G(s) = \int_0^1 \Delta_{\mathrm{DP}}(h_{t(\alpha)})\,d\alpha. \tag{8}$$

To measure predictive utility under the same deployment uncertainty, we define the *global accuracy* of s as

$$\mathrm{GAcc}(s) := \int_0^1 \mathrm{Acc}(h_{t(\alpha)})\,d\alpha. \tag{9}$$

Thus, rather than evaluating a score function at a single threshold, global fairness and global accuracy quantify its disparity and predictive utility, respectively, across possible acceptance rates. This motivates the population learning problem

$$\max_{s:\mathcal{X}\to\mathbb{R}} \quad \mathrm{GAcc}(s) \qquad \text{s.t.} \qquad \Delta_G(s) \leq \varepsilon. \tag{10}$$

This formulation seeks a single score function that supports flexible threshold selection at deployment without retraining, while explicitly accounting for how both fairness and accuracy vary across acceptance rates.

Estimating Global Fairness. To obtain a differentiable approximation suitable for gradient-based optimization, for $z \in \{-1, 1\}$, we define the smoothed empirical CDF $\widehat{F}_z(t) = \frac{1}{n_z}\sum_{i:Z_i=z} \Phi\Big(\frac{t-s(X_i)}{h}\Big)$, where $n_z := \sum_{i=1}^n \mathbf{1}\{Z_i = z\}$, $h > 0$ is a bandwidth parameter, and Φ is the Gaussian CDF. By definition Eq. (7), we estimate the global fairness gap by

$$\widehat{\Delta}_G(s) := \frac{1}{n}\sum_{i=1}^n \Big|\widehat{F}_1\big(s(X_i)\big) - \widehat{F}_{-1}\big(s(X_i)\big)\Big|. \tag{11}$$

Theorem 4.1 (Consistency of the empirical global fairness estimator)**.** *Let $\{(X_i, Z_i)\}_{i=1}^n$ be i.i.d. observations, where $Z_i \in \{-1, 1\}$. For each $z \in \{-1, 1\}$, let $\pi_z := \mathbb{P}(Z = z) > 0$ and $n_z := \sum_{i=1}^n \mathbf{1}\{Z_i = z\}$. Suppose that each F_z is absolutely continuous with density f_z satisfying $\|f_z\|_\infty \leq M < \infty$. Let $\widehat{\Delta}_G(s)$ be defined as Eq. (11). Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\Big|\widehat{\Delta}_G(s) - \Delta_G(s)\Big| \leq \sqrt{\frac{\log(6/\delta)}{2n_1}} + \sqrt{\frac{\log(6/\delta)}{2n_{-1}}} + \sqrt{\frac{\log(6/\delta)}{2n}} + 2Mh\sqrt{\frac{2}{\pi}}.$$

Moreover, as $n \to \infty$ and $h \to 0$, $|\widehat{\Delta}_G(s) - \Delta_G(s)| \xrightarrow{p} 0$. Therefore, the empirical global fairness estimator is consistent.

Theorem 4.1 establishes consistency with error $O_p(n^{-1/2} + h)$. Notably, the stochastic term retains the standard $n^{-1/2}$ rate, rather than the $(nh)^{-1/2}$ rate typical of kernel density estimation (Tsybakov, 2009), because our estimator smooths the empirical CDF rather than estimating the score density. The bandwidth h therefore contributes only to the smoothing bias, and consistency requires $n \to \infty$ and $h \to 0$, without the additional condition $nh \to \infty$.

Optimizing Global Accuracy. We first establish that global accuracy admits an equivalent characterization through AUC.

Theorem 4.2 (Global Accuracy and AUC)**.** *Assume that the distribution of $s(X)$ is continuous and that $0 < p_+ := \mathbb{P}(Y = 1) < 1$, with $p_- := \mathbb{P}(Y = -1) = 1 - p_+$. Let (X', Y', Z') be an independent copy of (X, Y, Z). Then the global accuracy satisfies*

$$\mathrm{GAcc}(s) = 2p_+p_- \,\mathrm{AUC}(s) + (0.5 - p_+p_-)\,, \tag{12}$$

where $\mathrm{AUC}(s) := \mathbb{P}(s(X) > s(X') \mid Y = 1,\ Y' = -1)$. Consequently, since $2p_+p_- > 0$, maximizing global accuracy is equivalent to maximizing AUC.

Theorem 4.2 allows us to optimize global accuracy using a pairwise AUC surrogate. We therefore optimize global accuracy using a pairwise AUC surrogate. Specifically, we adopt the pairwise logistic loss $\ell_\gamma(u) := \log\big(1 + \exp(-\gamma \cdot u)\big)$, where $\gamma > 0$ controls the smoothness, and minimize the resulting positive–negative pairwise ranking loss

$$L_{\mathrm{GAcc}}(s) := \frac{1}{n_+n_-} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}\{Y_i = 1\}\, \mathbf{1}\{Y_j = -1\}\, \ell_\gamma\big(s(X_i) - s(X_j)\big). \tag{13}$$

Efficient algorithms for AUC optimization have been extensively studied (Ying et al., 2016; Liu et al., 2020; Lei & Ying, 2021).

Training Objective and Optimization. Using Eq. (11) and Eq. (13), we obtain the following tractable empirical formulation of the fair learning problem Eq. (10):

$$\min_{s_\phi} \quad L_{\mathrm{GAcc}}(s_\phi) \qquad \text{s.t.} \qquad \widehat{\Delta}_G(s_\phi) \le \varepsilon, \tag{14}$$

where s_ϕ is a scoring model parameterized by ϕ. In our implementation, s_ϕ is instantiated as a neural network. To solve Eq. (14), we use the Lagrangian formulation

$$\min_{s_\phi} \max_{\lambda \ge 0} \quad L_{\mathrm{GAcc}}(s_\phi) + \lambda\big(\widehat{\Delta}_G(s_\phi) - \varepsilon\big), \tag{15}$$

where λ is the dual variable associated with the global fairness constraint. We optimize Eq. (15) by a primal–dual gradient method (Bertsekas, 1997): the model parameters ϕ are updated by gradient descent on the Lagrangian, while the dual variable λ is updated by projected gradient ascent.

5 Experiments

We evaluate the proposed framework from both global and conventional single-threshold perspectives. Our experiments are designed to answer three questions: *(i)* whether our method achieves stronger global accuracy–fairness trade-offs than existing threshold-level and score-level fairness methods; *(ii)* whether these improvements translate into fairness and predictive utility over a broader range of possible deployment acceptance rates; and *(iii)* whether optimizing global fairness preserves competitive performance under conventional single-threshold evaluation.

Datasets and baselines. We evaluate our method on three widely used fairness benchmarks: ACS (Ding et al., 2021), Adult (Becker & Kohavi, 1996), and COMPAS (Angwin et al., 2022). For ACS and Adult, the task is binary income prediction with sex as the sensitive attribute. For COMPAS, the task is two-year recidivism prediction with race as the sensitive attribute, specifically whether an individual is African-American.

We compare our method in Eq. (14) with empirical risk minimization (ERM), which does not impose a fairness constraint, and two categories of fair learning baselines: threshold-level methods, including Fairlearn (Agarwal et al., 2018), Minimax (Diana et al., 2021), and Fairret (Buyl et al., 2024); and score-level methods, including the Wasserstein-based method of Jiang et al. (2020) and the KS-based FairScore (Vogel et al., 2021). Additional experimental details are provided in Appendix C.

Global accuracy–fairness trade-offs. For each learned score function s_ε (or s_λ), obtained under a fairness budget ε (or regularization weight λ), we evaluate its behavior across target acceptance rates $\alpha \in \{0.1, \ldots, 0.9\}$. At each α, we choose the corresponding threshold $\theta_{\varepsilon,\alpha}$ such that $P_X(s_\varepsilon(X) \geq \theta_{\varepsilon,\alpha}) = \alpha$, inducing the classifier $h_{\varepsilon,\alpha}(x) := 2\mathbf{1}\{s_\varepsilon(x) \geq \theta_{\varepsilon,\alpha}\} - 1$. We compute the accuracy and DP/EO gap of each induced classifier and average these quantities across the evaluation acceptance rates to obtain empirical estimates of global accuracy $\mathrm{GAcc}(s)$ and the global fairness gap $\Delta_G(s)$. Each learned score function therefore contributes one point to Figures 2 and 7, while varying the fairness budget or regularization weight traces out the corresponding global accuracy–fairness trade-off. The algorithm and experimental results concerning EO are given in Appendix A.

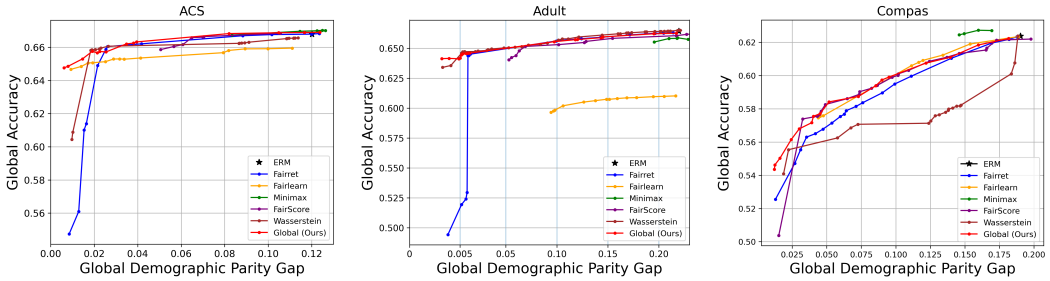


Figure 2: Global accuracy $\mathrm{GAcc}(s)$ versus global demographic parity gap $\Delta_G(s)$.

Figures 2 and 7 report the global accuracy–fairness trade-offs under DP and EO, respectively. Across the three datasets, our method generally achieves higher global accuracy at comparable levels of global fairness, with the advantage becoming particularly pronounced in the stringent-fairness regime. As the global fairness gap approaches zero, our method retains substantially more predictive utility, whereas competing methods often exhibit a sharper decrease in global accuracy. Importantly, the improvement is also observed over the KS- and Wasserstein-based score-level baselines, suggesting that how disparities across thresholds are aggregated matters for learning globally fair score functions. We next examine whether this improvement translates into the intended benefit across acceptance rates.

Fairness coverage across acceptance rates. The preceding results establish the aggregate global trade-off, but the central operational question is whether this improvement translates into desirable behavior over a broader range of possible deployment acceptance rates. Figures 3 and 4 therefore examine the induced classifiers at individual acceptance rates under comparable global performance levels.

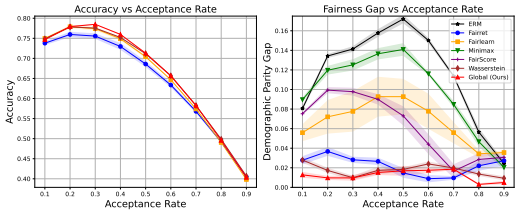


Figure 3: Accuracy and DP gap across acceptance rates on ACS. The left panel shows per-acceptance-rate accuracy for models with $\Delta_G(s) \leq 0.025$, and the right panel shows per-acceptance-rate DP gaps for models with $\mathrm{GAcc}(s) \geq 64.5\%$.

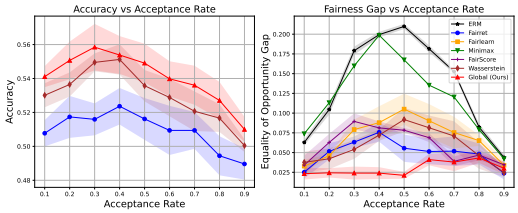


Figure 4: Accuracy and EO gap across acceptance rates on COMPAS. The left panel shows per-acceptance-rate accuracy for models with $\Delta_G(s) \leq 0.025$, and the right panel shows per-acceptance-rate EO gaps for models with $\mathrm{GAcc}(s) \geq 56.0\%$.

As shown in the right panel of Figure 3, among models with global accuracy of at least 64.5%, our method maintains a DP gap below 0.02 across the entire evaluated acceptance-rate range, whereas the competing methods satisfy this fairness level only over subsets of the range. Thus, at a comparable level of overall predictive utility, our learned score function supports a broader range of acceptance rates over which the induced classifiers satisfy the same fairness requirement. This provides greater flexibility when the deployment acceptance rate is unknown during training or needs to be adjusted after deployment.

This comparison is particularly informative relative to the KS- and Wasserstein-based score-level methods. Although both promote fairness beyond a single threshold, their objectives do not directly

characterize the range of population acceptance rates over which disparity persists. FairScore approximates KS-based distributional fairness through pointwise ROC-based fairness constraints, thereby emphasizing control of threshold-wise discrepancies rather than their coverage over population acceptance rates. Similarly, Wasserstein-1 integrates threshold-wise disparity uniformly over the raw score space, so equal-length score intervals receive equal weight regardless of the population mass they contain. Consequently, neither objective distinguishes whether a given disparity spans a broad or narrow range of acceptance rates. In contrast, global fairness weights threshold-wise disparity by the population score density, equivalently integrating it uniformly over population acceptance rates. The broader fair acceptance-rate coverage observed here is therefore consistent with the specific quantity targeted by our objective.

The left panel of Figure 3 provides the complementary comparison from the predictive-utility perspective. Among models with a global DP gap of at most 0.025, our method maintains higher accuracy across acceptance rates than the competing methods. Thus, the broader fairness coverage is not obtained simply by uniformly sacrificing predictive utility. The same qualitative behavior is observed under EO on COMPAS, as shown in Figure 4.

Standard single-threshold performance. We finally examine whether optimizing fairness globally across acceptance rates compromises performance under conventional single-threshold evaluation. Figures 5 and 8 report the corresponding accuracy–fairness Pareto frontiers. For each method, we vary the fairness budget ε or regularization weight λ and evaluate the resulting classifiers at the original decision thresholds and across different acceptance rates.

On ACS and Adult, the fairness-aware methods, including ours, achieve broadly similar accuracy–fairness frontiers and substantially improve over ERM. On COMPAS, our method achieves a stronger trade-off in the high-fairness regime while remaining comparable to the competing methods as the fairness requirement is relaxed. These results indicate that the improvements under global evaluation do not come at the expense of conventional operating-point performance: although our objective is designed to learn score functions with desirable fairness and utility across acceptance rates, the resulting models remain competitive when evaluated at individual thresholds.

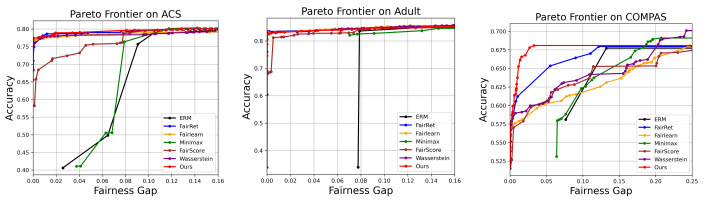


Figure 5: Accuracy versus DP gap under standard single-threshold evaluation.

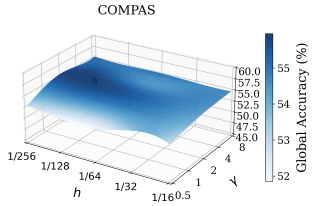


Figure 6: Parameter analysis of h and γ.

Parameter analysis. We analyze the sensitivity to the bandwidth h of the smoothed empirical CDF and the scale parameter γ of the pairwise accuracy surrogate on COMPAS under $\varepsilon = 0.3$. As shown in Figure 6, global accuracy is relatively stable across the considered bandwidths, whereas γ has a more noticeable effect, with a small value such as $\gamma = 0.5$ leading to lower performance. In our experiments, we fix $\gamma = 2$ and select h based on validation performance.

6 Conclusion

We studied fair score-function learning under acceptance-rate uncertainty, where the deployment threshold may be determined only after training. We introduced *global fairness*, which captures both the magnitude of group disparity and how broadly it persists across population acceptance rates. Its density-weighted formulation admits an equivalent interpretation as average disparity over acceptance rates and captures distinctions overlooked by KS and Wasserstein-1. We further introduced *global accuracy* and developed a learning framework that maximizes it subject to a global fairness constraint. Our theoretical results establish consistency of the empirical global fairness estimator and characterize global accuracy through its relationship with AUC. Experiments on three fairness benchmarks demonstrate favorable global accuracy–fairness trade-offs, broader fair acceptance-rate coverage, and competitive single-threshold performance. Overall, our framework provides a principled approach to learning fair and predictive score functions when the deployment acceptance rate is unknown in advance.

AI USE STATEMENT

In this work, we used generative AI tools to assist with intermediate mathematical derivations and provide ingredients for proving mathematical claims, assist with the writing and verification of proofs, provide feedback on research methodology and experimental design, and assist with the interpretation and presentation of experimental results. We did not use generative AI tools to generate synthetic datasets, develop the theoretical models or conceptual framework of the proposed method, formulate mathematical claims, propose or refine research hypotheses, implement the proposed method, assist with translation, or clean or reformat datasets. Qualitative and thematic data analysis was not applicable to this work.

Additionally, we used generative AI tools for brainstorming, summarizing and discussing relevant literature, suggesting experimental parameters, creating or editing auxiliary software code, drafting and restructuring parts of the manuscript, improving the clarity and readability of the writing, suggesting paper structure and presentation, and assisting with reference formatting.

All AI-assisted content was critically reviewed and revised by the authors. In particular, all AI-assisted mathematical derivations and proofs were independently checked by the authors, experimental choices and interpretations were validated against the actual experimental results, and AI-assisted code was reviewed and tested before use. The research questions, theoretical and conceptual framework, mathematical claims, final methodological choices, theoretical results, and experimental conclusions were determined and verified by the authors. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to support fairer automated decision-making under uncertainty in deployment acceptance rates. While the proposed framework may help reduce group disparities that arise when decision thresholds change after training, no single fairness metric can capture all relevant ethical considerations. The appropriate fairness notion and acceptable fairness–utility trade-off depend on the application context, affected populations, and potential harms. In high-stakes domains such as hiring, lending, and healthcare, the proposed method should therefore complement, rather than replace, domain-specific evaluation, transparency, and human oversight. Care should also be taken to avoid interpreting satisfaction of a particular fairness criterion as evidence that a decision system is fair in all respects. Our experiments use existing benchmark datasets and do not involve new data collection or human-subject experiments.

REPRODUCIBILITY STATEMENT

We provide the theoretical and experimental details needed to reproduce our results in the appendix and supplementary materials. Section A presents the extension of global fairness to equality of opportunity and the corresponding experimental results. Section B provides complete proofs of the theoretical results. Section C provides the experimental details required for reproduction, including the datasets, compared methods, hyperparameter settings, model architectures, and training configurations. We additionally provide the implementation code in the supplementary materials.

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.

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

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and machine learning: Limitations and opportunities. fairmlbook. org, 2019.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3): 334–334, 1997.

Maarten Buyl, MaryBeth Defrance, and Tijl De Bie. fairret: a framework for differentiable fairness regularization terms. In *The Twelfth International Conference on Learning Representations*, 2024.

Mingliang Chen and Min Wu. Towards threshold invariant fair classification. In *Conference on Uncertainty in Artificial Intelligence*, pp. 560–569. PMLR, 2020.

Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 66–76, 2021.

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.

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pp. 642–669, 1956.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Thomas J Espenshade, Chang Y Chung, and Joan L Walling. Admission preferences for minority students, athletes, and legacies at elite universities. *Social Science Quarterly*, 85(5):1422–1446, 2004.

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

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In *Uncertainty in artificial intelligence*, pp. 862–872. PMLR, 2020.

Kweku Kwegyir-Aggrey, Jessica Dai, A Feder Cooper, John Dickerson, Keegan Hines, and Suresh Venkatasubramanian. Repairing regressors for fair binary classification at any decision threshold. In *NeurIPS 2023 Workshop Optimal Transport and Machine Learning*, 2023.

Yunwen Lei and Yiming Ying. Stochastic proximal auc maximization. *Journal of Machine Learning Research*, 22(61):1–45, 2021. URL `http://jmlr.org/papers/v22/19-418.html`.

Mingrui Liu, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. Stochastic AUC maximization with deep neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pp. 1269–1283, 1990.

Margaret Sullivan Pepe. *The statistical evaluation of medical tests for classification and prediction*. Oxford university press, 2003.

Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 469–481, 2020.

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer, 2009.

Robin Vogel, Aurélien Bellet, and Stephan Clémençon. Learning fair scoring functions: Bipartite ranking under roc-based fairness constraints. In *International conference on artificial intelligence and statistics*, pp. 784–792. PMLR, 2021.

Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio P Calmon. Optimized score transformation for fair classification. In *AISTATS*, volume 20, pp. 1673–1683, 2020.

Yiming Ying, Longyin Wen, and Siwei Lyu. Stochastic online auc maximization. In *Advances in Neural Information Processing Systems (NIPS)*, volume 29, pp. 459–467, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/c52f1bd66cc19d05628bd8bf27af3ad6-Abstract.html.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180, 2017.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

This appendix provides additional theoretical and experimental details supporting the main paper. Section A presents the extension of global fairness to equality of opportunity and the corresponding experimental results. Section B provides complete proofs of the theoretical results, and Section C describes the datasets, compared methods, and experimental configurations used in our evaluation.

A Extension of the Global Fairness Gap to Equality of Opportunity

The equality-of-opportunity case admits a parallel characterization, obtained by replacing the group-wise score CDFs with the corresponding class-conditional score CDFs among positively labeled samples. When the fairness notion is equality of opportunity (EO), the disparity function of the score s is defined as

$$D_s(t) := |P(s(X) \geq t \mid Z = 1, Y = 1) - P(s(X) \geq t \mid Z = -1, Y = 1)| = |F_{1,+}(t) - F_{-1,+}(t)|,$$

where the group-wise CDF in the positive class is defined as

$$F_{z,+}(t) = \mathbb{P}(s(X) \leq t \mid Z = z, Y = 1),$$

for $z \in \{-1, 1\}$. Then, similar to Eq. (7), we define the EO-based global fairness as

$$\Delta_G(s) := \int_{\mathcal{T}} D_s(t) f(t)\, dt, \tag{16}$$

where f denotes the density of the population score distribution and $\mathcal{T}$ its support.

For $z \in \{-1, 1\}$, define the smoothed CDF estimator

$$\widehat{F}_{z,+}(t) = \frac{1}{n_{z,+}} \sum_{i:Z_i=z,Y_i=1} \Phi\left(\frac{t - s(X_i)}{h}\right),$$

where $n_{z,+} := \sum_{i=1}^{n} \mathbf{1}\{Z_i = z, Y_i = 1\}$, $h > 0$ is a bandwidth parameter, and Φ is the Gaussian CDF.

By definition Eq. (16), we estimate the global fairness gap by

$$\widehat{\Delta}_G(s) := \frac{1}{n} \sum_{i=1}^{n} \left| \widehat{F}_{1,+}\big(s(X_i)\big) - \widehat{F}_{-1,+}\big(s(X_i)\big) \right|. \tag{17}$$

Then we follow Eq. (15) to solve the score function.

We report the global accuracy–fairness trade-off and single-threshold performance under EO below.

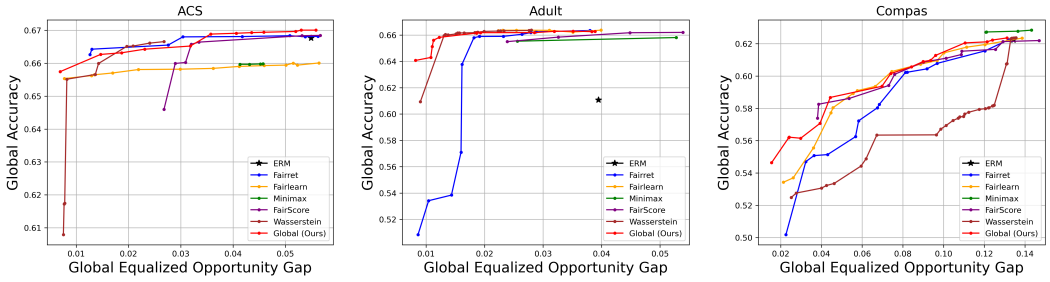


Figure 7: Global accuracy $\mathrm{GAcc}(s)$ vs. global equality-of-opportunity gap $\Delta_G(s)$ on datasets.

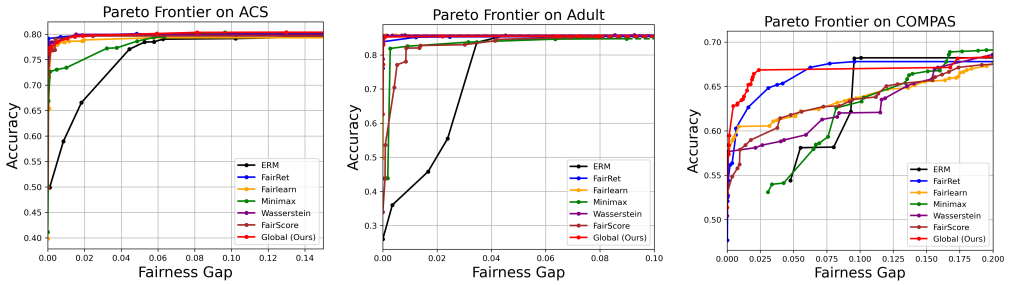
B Proofs

B.1 Proofs Related to Section 3

Proof of Theorem 3.3. Since $\alpha(t) = 1 - F_s(t)$, we have $d\alpha = -f_s(t)\, dt$. Applying the change of variables $\alpha = 1 - F_s(t)$ to Definition 3.2 gives

$$\Delta_G(s) = \int_{-\infty}^{+\infty} D_s(t) f_s(t)\, dt = \int_1^0 D_s(t(\alpha))(-d\alpha) = \int_0^1 D_s\big(t(\alpha)\big)\, d\alpha,$$

which yields the conclusion. □



(a) ACS (b) Adult (c) COMPAS

Figure 8: Accuracy versus EO gap under the standard single-operating-point evaluation.

Proof of Example 3.1. For any $c \in [30, 70]$, the two group-conditional distributions share the components $0.8\mathcal{N}(30, 9)$ and $0.1\mathcal{N}(90, 9)$. These common components therefore cancel when taking the difference between the two group-wise CDFs. Let $F_{1,c}$ and $F_{-1,c}$ denote the cumulative distribution functions of the score $s_c(X)$ conditional on $A = 1$ and $A = -1$, respectively. Hence,

$$\begin{aligned} D_c(t) &= |F_{1,c}(t) - F_{-1,c}(t)| \\ &= 0.1 \left| \Phi\left(\frac{t-(c-10)}{3}\right) - \Phi\left(\frac{t-(c+10)}{3}\right) \right|. \end{aligned}$$

Since $c - 10 < c + 10$,

$$D_c(t) = 0.1 \left[\Phi\left(\frac{t-(c-10)}{3}\right) - \Phi\left(\frac{t-(c+10)}{3}\right) \right].$$

Let

$$D_0(u) = 0.1 \left[\Phi\left(\frac{u+10}{3}\right) - \Phi\left(\frac{u-10}{3}\right) \right].$$

Then

$$D_c(t) = D_0(t - c).$$

Thus, changing c only translates the disparity function along the score axis without changing either its shape, maximum magnitude, or area.

The maximum of $D_c(t)$ is attained at its midpoint $t = c$. Therefore,

$$\begin{aligned} \mathrm{KS}(s_c) &= D_c(c) \\ &= 0.1 \left[\Phi\left(\frac{10}{3}\right) - \Phi\left(-\frac{10}{3}\right) \right] \\ &= 0.1 \left[2\Phi\left(\frac{10}{3}\right) - 1 \right]. \end{aligned}$$

Hence,

$$\mathrm{KS}(s_c) = 0.1 \left[2\Phi\left(\frac{10}{3}\right) - 1 \right]$$

for every $c \in [30, 70]$.

For the Wasserstein distance,

$$\begin{aligned} W_1(s_c) &= \int_{\mathbb{R}} D_c(t)\, dt \\ &= 0.1 \int_{\mathbb{R}} \left[\Phi\left(\frac{t-(c-10)}{3}\right) - \Phi\left(\frac{t-(c+10)}{3}\right) \right] dt. \end{aligned}$$

Using

$$\int_{\mathbb{R}} \left[\Phi\left(\frac{t-\mu_1}{\sigma}\right) - \Phi\left(\frac{t-\mu_2}{\sigma}\right) \right] dt = \mu_2 - \mu_1$$

for $\mu_1 < \mu_2$, we obtain

$$W_1(s_c) = 0.1\big[(c+10) - (c-10)\big] = 2.$$

Therefore,

$$W_1(s_c) = 2$$

for every $c \in [30, 70]$.

We next analyze global fairness. Since the two sensitive groups are equally likely, the marginal score density induced by s_c is

$$f_{s,c}(t) = 0.05\phi_{c-10,3}(t) + 0.8\phi_{30,3}(t) + 0.05\phi_{c+10,3}(t) + 0.1\phi_{90,3}(t),$$

where $\phi_{\mu,\sigma}$ denotes the density of $\mathcal{N}(\mu, \sigma^2)$. Therefore,

$$\Delta_G(s_c) = \int_{\mathbb{R}} D_c(t) f_{s,c}(t)\, dt.$$

For a Gaussian random variable $T \sim \mathcal{N}(\mu, 9)$, the Gaussian convolution identity gives

$$\mathbb{E}\left[\Phi\left(\frac{T-a}{3}\right)\right] = \Phi\left(\frac{\mu-a}{3\sqrt{2}}\right).$$

Consequently,

$$\mathbb{E}\left[D_c(T)\right] = 0.1\left[\Phi\left(\frac{\mu-(c-10)}{3\sqrt{2}}\right) - \Phi\left(\frac{\mu-(c+10)}{3\sqrt{2}}\right)\right].$$

Applying this identity to the four Gaussian components of $f_{s,c}$ yields

$$\begin{aligned}\Delta_G(s_c) =& 0.005\left[\Phi(0) - \Phi\left(-\frac{20}{3\sqrt{2}}\right)\right] \\ &+ 0.08\left[\Phi\left(\frac{40-c}{3\sqrt{2}}\right) - \Phi\left(\frac{20-c}{3\sqrt{2}}\right)\right] \\ &+ 0.005\left[\Phi\left(\frac{20}{3\sqrt{2}}\right) - \Phi(0)\right] \\ &+ 0.01\left[\Phi\left(\frac{100-c}{3\sqrt{2}}\right) - \Phi\left(\frac{80-c}{3\sqrt{2}}\right)\right].\end{aligned}$$

Using $\Phi(-x) = 1 - \Phi(x)$ and $\Phi(0) = 1/2$, the first and third terms combine, giving

$$\begin{aligned}\Delta_G(s_c) =& 0.01\left[\Phi\left(\frac{20}{3\sqrt{2}}\right) - \frac{1}{2}\right] \\ &+ 0.08\left[\Phi\left(\frac{40-c}{3\sqrt{2}}\right) - \Phi\left(\frac{20-c}{3\sqrt{2}}\right)\right] \\ &+ 0.01\left[\Phi\left(\frac{100-c}{3\sqrt{2}}\right) - \Phi\left(\frac{80-c}{3\sqrt{2}}\right)\right].\end{aligned}$$

Unlike KS and Wasserstein fairness, this quantity explicitly depends on c. At $c = 30$, the disparity function is centered at score 30, exactly where the dominant 80% population component is concentrated. Substitution gives

$$\Delta_G(f_{30}) \approx 0.0835.$$

At $c = 70$, the same disparity shape has been translated into a substantially less populated region of the score distribution, giving

$$\Delta_G(f_{70}) \approx 0.0051.$$

Hence,

$$\frac{\Delta_G(f_{30})}{\Delta_G(f_{70})} \approx 16.4.$$

Therefore, the entire classifier family has identical KS and Wasserstein fairness, even though the location of the disparity relative to the population distribution changes continuously with c. Global fairness captures precisely this population dependence through the weighting density $f_{s,c}(t)$, and therefore distinguishes classifiers whose disparity affects substantially different portions of the population. □

B.2 Proofs Related to Section 4

Proof of Theorem 4.1. For notational simplicity, define

$$g_s(x) := \left|F_1\big(s(x)\big) - F_{-1}\big(s(x)\big)\right|,$$

and

$$\widehat{g}_s^h(x) := \left|\widehat{F}_1^h\big(s(x)\big) - \widehat{F}_{-1}^h\big(s(x)\big)\right|.$$

Let $\mathbb{P}$ denote expectation with respect to X, and let $\mathbb{P}_n$ denote the empirical measure

$$\mathbb{P}_n\varphi := \frac{1}{n}\sum_{i=1}^{n}\varphi(X_i).$$

By definition,

$$\Delta_G(s) = \mathbb{P}g_s$$

and

$$\widehat{\Delta}_G(s) = \mathbb{P}_n\widehat{g}_s^h.$$

Therefore,

$$\begin{aligned}\left|\widehat{\Delta}_G(s) - \Delta_G(s)\right| &= \left|\mathbb{P}_n\widehat{g}_s^h - \mathbb{P}g_s\right| \\ &\le \left|\mathbb{P}_n\widehat{g}_s^h - \mathbb{P}_n g_s\right| + \left|\mathbb{P}_n g_s - \mathbb{P}g_s\right| \\ &\le \|\widehat{g}_s^h - g_s\|_\infty + \left|(\mathbb{P}_n - \mathbb{P})g_s\right|.\end{aligned}$$

This deterministic decomposition does not require the quantities $\widehat{g}_s^h(X_i)$ to be independent across i. In particular, it remains valid even though the same observations are used both to construct $\widehat{F}_z^h$ and to evaluate the outer empirical average.

We first control $\|\widehat{g}_s^h - g_s\|_\infty$. Since the absolute value function is 1-Lipschitz, for every $x \in \mathcal{X}$,

$$\begin{aligned}\left|\widehat{g}_s^h(x) - g_s(x)\right| \le &\left|\widehat{F}_1^h\big(s(x)\big) - F_1\big(s(x)\big)\right| \\ &+ \left|\widehat{F}_{-1}^h\big(s(x)\big) - F_{-1}\big(s(x)\big)\right|.\end{aligned}$$

Hence,

$$\|\widehat{g}_s^h - g_s\|_\infty \le \|\widehat{F}_1^h - F_1\|_\infty + \|\widehat{F}_{-1}^h - F_{-1}\|_\infty.$$

We next bound each smoothed empirical CDF error. Define the ordinary empirical CDF within group z by

$$\widetilde{F}_z(t) := \frac{1}{n_z}\sum_{i:Z_i=z}\mathbf{1}\{s(X_i) \le t\},$$

and define the population-smoothed CDF by

$$F_z^h(t) := \mathbb{E}\left[\Phi\left(\frac{t - s(X)}{h}\right)\,\middle|\, Z = z\right].$$

By the triangle inequality,

$$\|\widehat{F}_z^h - F_z\|_\infty \le \|\widehat{F}_z^h - F_z^h\|_\infty + \|F_z^h - F_z\|_\infty.$$

Let $\xi \sim \mathcal{N}(0,1)$ be independent of the observations. Since

$$\Phi\left(\frac{t-u}{h}\right) = \mathbb{P}_\xi(u + h\xi \le t),$$

we have

$$\widehat{F}_z^h(t) = \mathbb{E}_\xi\left[\widetilde{F}_z(t - h\xi)\right]$$

and

$$F_z^h(t) = \mathbb{E}_\xi\left[F_z(t - h\xi)\right].$$

Consequently,

$$\begin{aligned}\left|\widehat{F}_z^h(t) - F_z^h(t)\right| &= \left|\mathbb{E}_\xi\left[\widetilde{F}_z(t-h\xi) - F_z(t-h\xi)\right]\right| \\ &\le \mathbb{E}_\xi\left[\left|\widetilde{F}_z(t-h\xi) - F_z(t-h\xi)\right|\right] \\ &\le \|\widetilde{F}_z - F_z\|_\infty.\end{aligned}$$

Taking the supremum over t gives

$$\|\widehat{F}_z^h - F_z^h\|_\infty \le \|\widetilde{F}_z - F_z\|_\infty.$$

Conditional on the group labels and hence on n_z, the observations $\{s(X_i) : Z_i = z\}$ are i.i.d. with CDF F_z. The Dvoretzky–Kiefer–Wolfowitz inequality with Massart's sharp constant (Dvoretzky et al., 1956; Massart, 1990) gives

$$\mathbb{P}\left(\|\widetilde{F}_z - F_z\|_\infty > \epsilon \;\middle|\; n_z\right) \le 2\exp(-2n_z\epsilon^2).$$

Therefore, for any $\delta_z \in (0,1)$, with conditional probability at least $1-\delta_z$,

$$\|\widehat{F}_z^h - F_z^h\|_\infty \le \sqrt{\frac{\log(2/\delta_z)}{2n_z}}.$$

It remains to control the smoothing bias. Since $\|f_z\|_\infty \le M$, each CDF F_z is M-Lipschitz. Therefore, for every $t \in \mathbb{R}$,

$$\begin{aligned}\left|F_z^h(t) - F_z(t)\right| &= \left|\mathbb{E}_\xi\left[F_z(t-h\xi) - F_z(t)\right]\right| \\ &\le \mathbb{E}_\xi\left[\left|F_z(t-h\xi) - F_z(t)\right|\right] \\ &\le Mh\mathbb{E}|\xi| \\ &= Mh\sqrt{\frac{2}{\pi}}.\end{aligned}$$

Hence,

$$\|F_z^h - F_z\|_\infty \le Mh\sqrt{\frac{2}{\pi}},$$

and thus, with probability at least $1-\delta_z$,

$$\|\widehat{F}_z^h - F_z\|_\infty \le \sqrt{\frac{\log(2/\delta_z)}{2n_z}} + Mh\sqrt{\frac{2}{\pi}}.$$

Combining the bounds for the two sensitive groups gives

$$\begin{aligned}\|\widehat{g}_s^h - g_s\|_\infty \le &\sqrt{\frac{\log(2/\delta_1)}{2n_1}} + \sqrt{\frac{\log(2/\delta_{-1})}{2n_{-1}}} \\ &+ 2Mh\sqrt{\frac{2}{\pi}}.\end{aligned}$$

Finally, since F_1 and F_{-1} take values in $[0,1]$,

$$0 \le g_s(x) \le 1.$$

Hence, Hoeffding's inequality for bounded independent random variables (Hoeffding, 1963) implies that, for any $\delta_0 \in (0,1)$, with probability at least $1-\delta_0$,

$$|(\mathbb{P}_n - \mathbb{P})g_s| \le \sqrt{\frac{\log(2/\delta_0)}{2n}}.$$

Taking $\delta_0 = \delta_1 = \delta_{-1} = \delta/3$ and applying the union bound yields, with probability at least $1-\delta$,

$$\left|\widehat{\Delta}_G(s) - \Delta_G(s)\right| \le \sqrt{\frac{\log(6/\delta)}{2n_1}} + \sqrt{\frac{\log(6/\delta)}{2n_{-1}}} + \sqrt{\frac{\log(6/\delta)}{2n}} + 2Mh\sqrt{\frac{2}{\pi}}.$$

Finally, since

$$\frac{n_z}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Z_i = z\},$$

the strong law of large numbers gives

$$\frac{n_z}{n} \xrightarrow{\text{a.s.}} \pi_z > 0$$

for each $z \in \{-1, 1\}$. Hence, $n_z = \Theta_p(n)$ and

$$\sqrt{\frac{1}{n_1}} + \sqrt{\frac{1}{n_{-1}}} + \sqrt{\frac{1}{n}} = O_p(n^{-1/2}).$$

Combining this with the preceding bound yields

$$\left|\widehat{\Delta}_G^{h_n}(s) - \Delta_G(s)\right| = O_p\left(n^{-1/2} + h_n\right).$$

Therefore, as $n \to \infty$ and $h_n \to 0$,

$$\left|\widehat{\Delta}_G^{h_n}(s) - \Delta_G(s)\right| \xrightarrow{p} 0,$$

which proves

$$\widehat{\Delta}_G^{h_n}(s) \xrightarrow{p} \Delta_G(s).$$

□

Proof of Theorem 4.2. Let

$$F_+(t) := \mathbb{P}(s(X) \le t \mid Y = 1), \qquad F_-(t) := \mathbb{P}(s(X) \le t \mid Y = -1),$$

and let

$$F(t) := \mathbb{P}(s(X) \le t) = p_+ F_+(t) + p_- F_-(t).$$

For an acceptance rate $\alpha \in [0, 1]$, let $\theta_\alpha := F^{-1}(1 - \alpha)$ denote the corresponding score threshold. Since the distribution of $s(X)$ is continuous,

$$\alpha = \mathbb{P}(s(X) > \theta_\alpha) = 1 - F(\theta_\alpha).$$

By the definition of the thresholded classifier and the law of total probability,

$$\begin{aligned}\mathrm{Acc}(f_\alpha) &= p_+ \,\mathbb{P}(s(X) > \theta_\alpha \mid Y = 1) + p_- \,\mathbb{P}(s(X) \le \theta_\alpha \mid Y = -1) \\ &= p_+\big(1 - F_+(\theta_\alpha)\big) + p_- F_-(\theta_\alpha). \end{aligned} \tag{18}$$

Using $p_- F_-(t) = F(t) - p_+ F_+(t)$, we have

$$p_+\big(1 - F_+(t)\big) + p_- F_-(t) = F(t) - p_+ + 2p_+\big(1 - F_+(t)\big). \tag{19}$$

Therefore, using the change of variable $\alpha = 1 - F(t)$,

$$\begin{aligned}\mathrm{GAcc}(s) &= \int_0^1 \mathrm{Acc}(f_\alpha)\, d\alpha \\ &= \int_{-\infty}^{\infty} \big[F(t) - p_+ + 2p_+\big(1 - F_+(t)\big)\big]\; dF(t). \end{aligned} \tag{20}$$

The first term satisfies

$$\int_{-\infty}^{\infty} \big(F(t) - p_+\big)\, dF(t) = \frac{1}{2} - p_+, \tag{21}$$

because $F(s(X))$ is uniformly distributed on $[0, 1]$ under continuity.

For the remaining term, let (X', Y', Z') be an independent copy of (X, Y, Z). Then

$$\int_{-\infty}^{\infty} \big(1 - F_+(t)\big)\, dF(t) = \int_{-\infty}^{\infty} \mathbb{P}(s(X) > t \mid Y = 1)\, dF(t) = \mathbb{P}\big(s(X) > s(X') \mid Y = 1\big). \tag{22}$$

Combining Eq. (20)–Eq. (22) gives

$$\mathrm{GAcc}(s) = \frac{1}{2} - p_+ + 2p_+\, \mathbb{P}\big(s(X) > s(X') \mid Y = 1\big). \tag{23}$$

Applying the law of total probability with respect to Y',

$$\mathbb{P}\big(s(X) > s(X') \mid Y = 1\big) = p_+\, \mathbb{P}\big(s(X) > s(X') \mid Y = 1, Y' = 1\big) + p_-\, \mathbb{P}\big(s(X) > s(X') \mid Y = 1, Y' = -1\big). \tag{24}$$

Conditional on $Y = Y' = 1$, the random variables $s(X)$ and $s(X')$ are independent and identically distributed. Since their distribution is continuous, ties occur with probability zero, and symmetry implies

$$\mathbb{P}\big(s(X) > s(X') \mid Y = 1, Y' = 1\big) = \frac{1}{2}.$$

Moreover, by definition,

$$\mathbb{P}\big(s(X) > s(X') \mid Y = 1, Y' = -1\big) = \mathrm{AUC}(s).$$

Hence,

$$\mathbb{P}\big(s(X) > s(X') \mid Y = 1\big) = \frac{p_+}{2} + p_-\, \mathrm{AUC}(s). \tag{25}$$

Substituting Eq. (25) into Eq. (23), we obtain

$$\begin{aligned} \mathrm{GAcc}(s) &= \frac{1}{2} - p_+ + 2p_+ \left(\frac{p_+}{2} + p_-\, \mathrm{AUC}(s)\right) \\ &= \frac{1}{2} - p_+ + p_+^2 + 2p_+p_-\, \mathrm{AUC}(s) \\ &= \frac{1}{2} - p_+p_- + 2p_+p_-\, \mathrm{AUC}(s), \end{aligned} \tag{26}$$

where the last equality uses $p_- = 1 - p_+$.

Since $0 < p_+ < 1$, we have $2p_+p_- > 0$. Therefore, $\mathrm{GAcc}(s)$ is a strictly increasing affine function of $\mathrm{AUC}(s)$, and maximizing global accuracy is equivalent to maximizing AUC. □

C Experimental Details

C.1 Datasets

ACSIncome. The ACSIncome dataset is derived from the American Community Survey using the Folktables framework. The task is to predict whether an individual's annual income exceeds a threshold (binary classification), and the sensitive attribute is sex (male vs. female). Compared with traditional fairness benchmarks, this dataset is large-scale and reflects more realistic population distributions.

Adult. The Adult dataset (also known as Census Income) is a classic benchmark for fairness research. The task is to predict whether an individual's annual income exceeds \$50k, and the sensitive attribute is sex. It is a relatively small tabular dataset and is widely used for evaluating fairness-aware algorithms.

COMPAS. The COMPAS dataset is used for recidivism prediction, where the task is to predict whether a defendant will reoffend within two years. The sensitive attribute is race, defined as whether the individual is African-American. This dataset is commonly used to study fairness issues in high-stakes decision-making systems.

C.2 Compared Methods

ERM. Empirical risk minimization (ERM) solves the classification problem by minimizing the cross-entropy loss without fairness constraints, i.e.,

$$h := \arg\min_h \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, h(X_i)).$$

Fairret. Fairret learns a classifier by optimizing a linear combination of prediction error and a differentiable fairness-gap penalty:

$$h_\lambda := \arg\min_h \ \mathrm{Err}(h) + \lambda \cdot \Delta(h),$$

where $\lambda > 0$ controls the fairness regularization strength. Varying λ produces a family of classifiers h_λ with corresponding logit score functions s_λ satisfying $h_\lambda(x) = 2\mathbf{1}\{s_\lambda(x) \geq 0\} - 1$. The grid of the fairness strength λ is $\{0.005, 0.01, 0.1, 0.5, 1, 5, 10, 20, 50, 100\}$.

Minimax. Minimax models fairness through a minimax formulation, where the learner minimizes the prediction loss while controlling the maximum disparity across groups. In our experiments, we use the objective

$$h_\varepsilon := \arg\min_h P(h(X) \neq Y)$$
$$\text{such that } \max_{z \in \mathcal{Z}} P(h(X) \neq Y \mid Z = z) \leq \varepsilon,$$

where $\varepsilon > 0$ controls the maximum group error budget. This formulation enforces a worst-group guarantee and yields a family of models indexed by ε. The grid of ε is $\{0.01, 0.02, 0.03, 0.05, 0.1, 0.15, 0.2, 0.3\}$.

Fairlearn. Fairlearn considers the following constrained optimization problem over randomized classifiers:

$$Q^*_\varepsilon := \arg\min_{Q \in \Delta} \mathrm{Err}(Q)$$
$$\text{such that} \quad M\mu(Q) \leq \varepsilon,$$

where Δ is the set of distributions over the classifier space $\mathcal{F}$, the matrix M and vector ε specify the linear fairness constraints, and $\mu(Q) := \sum_{h \in \mathcal{F}} Q(h)\mu(h)$, where $\mu(h)$ is the corresponding vector of error terms. Fairlearn reduces the problem to weighted classification and solves it via exponentiated gradient, yielding a mixture of classifiers

$$h_\varepsilon = \sum_t w_t h_{t,\varepsilon},$$

where w_t is the mixture weight and $h_{t,\varepsilon}$ is a base classifier. Our global evaluation protocol requires a score/logit function that can be thresholded to realize varying acceptance rates. For Fairlearn, in addition to the score function of the mixture classifiers h_ε, we also use the score functions of all base classifiers $h_{t,\varepsilon}$ to compute the global accuracy and fairness gaps. During implementation, the grid of ε is $\{0.001, 0.005, 0.01, 0.02, 0.05\}$.

FairScore. FairScore learns a scoring function through bipartite ranking while promoting fairness via pointwise ROC-based constraints. Following the original setup, we sweep the shared fairness weight over $\lambda \in \{0, 0.25, 0.5, 1, 5, 10\}$ to obtain different accuracy–fairness trade-offs. For each λ, the ℓ_2-regularization parameter is selected on the validation set from $\lambda_{\mathrm{reg}} \in \{10^{-3}, 5 \times 10^{-3}, 10^{-2}, 5 \times 10^{-2}, 10^{-1}, 5 \times 10^{-1}, 1\}$.

Wasserstein. The Wasserstein-based method promotes score-level fairness by penalizing the Wasserstein-1 discrepancy between sensitive-group score distributions. We implement its empirical objective using a differentiable Wasserstein-1 surrogate and sweep the classification–fairness trade-off parameter over $\alpha \in \{0.99, 0.95, 0.9, 0.8, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1\}$ and the fairness penalty weight over $\beta \in \{0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100, 200, 500\}$ to trace the accuracy–fairness trade-off. The Wasserstein term is weighted by $(1 - \alpha)\beta$, while the classification loss is weighted by α. We use a Pseudo-Huber approximation with smoothing parameter $\delta = 0.1$ and optimize the model using Adam with learning rate 10^{-3}.

C.3 Implementation Details

We implement all baseline methods using their publicly available official codebases. All experiments are implemented in PyTorch on a workstation equipped with an Intel(R) Core(TM) i7-14700K CPU @ 3.40GHz and an NVIDIA RTX A5000 GPU with 24GB memory. All methods use a two-layer ReLU neural network with hidden dimension 64 as the scoring model, and we adopt a standard $60\%/20\%/20\%$ train-validation-test split.

For our method, the learning rate is set to 10^{-3}, the scale parameter is fixed at 2, and the kernel bandwidth is selected from the grid $\{2^{-4}, 2^{-5}, \ldots, 2^{-8}\}$. The range of our fairness budget ε is $\{0.005, 0.010, 0.015, 0.020, \ldots, 0.10, 0.11, 0.12, \ldots, \Delta_G(h_{\mathrm{ERM}})\}$.