FairPOT: Balancing AUC Performance and Fairness with Proportional Optimal Transport

Pengxi Liu, Yi Shen, Matthew M. Engelhard, Benjamin A. Goldstein, Michael J. Pencina, Nicoleta J. Economou-Zavlanos, Michael M. Zavlanos

Duke University

pengxi.liu@duke.edu, yi.shen478@duke.edu, m.engelhard@duke.edu, ben.goldstein@duke.edu, michal.pencina@duke.edu, nicoleta.economou@duke.edu, mz61@duke.edu

Abstract

Fairness metrics utilizing the area under the receiver operator characteristic curve (AUC) have gained increasing attention in high-stakes domains such as healthcare, finance, and criminal justice. In these domains, fairness is often evaluated over risk scores rather than binary outcomes, and a common challenge is that enforcing strict fairness can significantly degrade AUC performance. To address this challenge, we propose Fair Proportional Optimal Transport (FairPOT), a novel, model-agnostic post-processing framework that strategically aligns risk score distributions across different groups using optimal transport, but does so selectively by transforming a controllable proportion, i.e., the top-λ quantile, of scores within the disadvantaged group. By varying λ, our method allows for a tunable trade-off between reducing AUC disparities and maintaining overall AUC performance. Furthermore, we extend FairPOT to the partial AUC setting, enabling fairness interventions to concentrate on the highest-risk regions. Extensive experiments on synthetic, public, and clinical datasets show that FairPOT consistently outperforms existing post-processing techniques in both global and partial AUC scenarios, often achieving improved fairness with slight AUC degradation or even positive gains in utility. The computational efficiency and practical adaptability of FairPOT make it a promising solution for real-world deployment.

Introduction

Recent advancements in machine learning have raised equal excitement and concern, especially in high-stakes domains that involve a large extent of decision-making, such as healthcare (Rajkomar et al. 2018), finance (Hardt, Price, and Srebro 2016), and criminal justice (Ensign et al. 2018). A major concern lies in the biases embedded throughout the machine learning lifecycle (Suresh and Guttag 2021). These biases may arise from various stages, including data collection, algorithm design, and model deployment, and can result in systematic discrimination against certain individuals or demographic groups (Mehrabi et al. 2021). In this sense, fairness has emerged as a prominent topic in the study of machine learning ethics. Within this realm, researchers aim to address two fundamental questions: *conceptually*, how to define and measure fairness, and *methodologically*, how to achieve fairness (Pessach and Shmueli 2022).

A widely studied class of fairness notions is group fairness (Mehrabi et al. 2021). It compares the outcomes of machine learning models across different groups defined by sensitive attributes, such as gender or race (Caton and Haas 2024). Classical group fairness notions, such as demographic parity (Dwork et al. 2012) and equalized odds (Hardt, Price, and Srebro 2016), are framed around binary outcomes. They are typically evaluated using confusion matrices or statistical parity after thresholding continuous risk scores into binary predictions (Kallus and Zhou 2019; Caton and Haas 2024).

However, in many real-world applications, especially in high-stakes settings, risk scores, i.e., continuous-valued outputs, are used directly to guide decisions. Risk scores convey the degree of uncertainty in model predictions and serve as intermediate outputs, rather than binary decisions. Unlike deterministic binary outcomes, unadjusted risk scores allow for more flexible and context-aware interventions (Kallus and Zhou 2019). For instance, in healthcare, a patient with a predicted risk score of 0.7 may receive priority compared to a patient with a risk score of 0.6, even if both are above a decision threshold. In this context, fairness concerns arise at the level of the scores themselves, motivating the need for methods that can address score-level group disparities, rather than just binary outcomes.

To this end, we focus on score-based group fairness. A representative metric in this setting is xAUC disparity (Kallus and Zhou 2019), which measures fairness from a bipartite ranking perspective. Specifically, xAUC measures the probability that risk scores rank randomly chosen positive samples from one group higher than the negative samples from another group, and vice versa. The xAUC disparity is then defined as the difference between these two probabilities, capturing the asymmetry in class distinguishability across groups.

To mitigate score-level disparities, various fairness-enhancing strategies have been proposed. Broadly, these can be categorized into pre-processing (modifying the input data), in-processing (changing the learning algorithm), and post-processing (adjusting model outputs) approaches in terms of the underlying stage at which fairness interventions are implemented (Pessach and Shmueli 2022; Caton and Haas 2024). In this work, we focus on post-processing methods, which are particularly appealing for their model-agnostic nature: they can be applied to any predictive model without modifying the training process or requiring access to

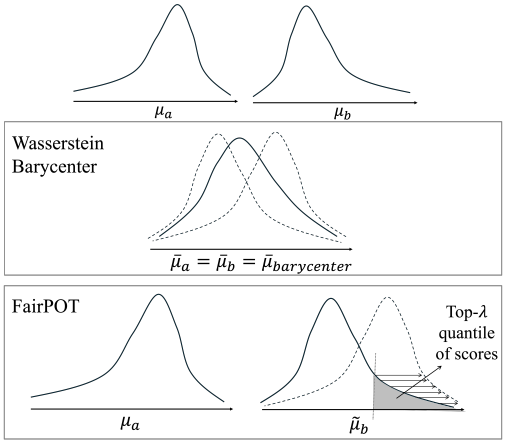

Figure 1: Comparison of FairPOT and Wasserstein barycenter-based methods.

internal parameters. This makes post-processing especially practical in real-world settings where models are already deployed or not easily modifiable.

Among post-processing techniques, optimal transport has gained increasing attention as it provides a principled foundation for distribution alignment. Existing studies typically move score distributions from different groups toward a shared target, named as the Wasserstein barycenter (Jiang et al. 2020; Silvia et al. 2020; Xian, Yin, and Zhao 2023). As such, the post-processed scores follow identical distributions, reducing group-level discrepancies and promoting fairness at the score level. However, such repairs can introduce substantial deviations from the original scores, potentially harming model utility.

In this work, we propose *Fair Proportional Optimal Transport* (FairPOT), which, unlike related literature, applies optimal transport only to a proportion (e.g., top-λ quantile) of the disadvantaged group, while keeping the advantaged group unchanged, as illustrated in Figure 1. By limiting the adjustment to a selected portion of scores from the disadvantaged group, FairPOT ensures no modification to the advantaged group and provides precise control over the extent of alignment applied to the disadvantaged group. The parameter λ allows for balancing fairness and accuracy—in our case, xAUC disparity and AUC. We further extend FairPOT to the partial AUC (pAUC) setting, where fairness and accuracy are evaluated within top-ranked regions of the score distribution, which are most critical in high-stakes decision-making scenarios. For instance, in medical screening tasks where the overall positive rate is low, it is often more important to ensure that high-risk individuals (i.e., those assigned high risk scores) are correctly prioritized, rather than focusing on the entire population. Our proposed method enables explicit balancing between fairness and accuracy in both global (AUC) and partial (pAUC) settings. We characterize the Pareto frontier of the fairness–accuracy plot as the set of solutions where no further improvement in fairness can be achieved without sacrificing accuracy, and vice versa. We validate our approach through extensive experiments on a variety of datasets, including synthetic, public, and real-world clinical data.

Related Work

AUC-based Fairness

AUC-based fairness metrics are particularly important for evaluating fairness of risk scores, as they assess ranking quality independently of thresholding. One early metric is pinned AUC (Dixon et al. 2018), which evaluates AUC on balanced datasets with equal subgroup representation. Borkan et al. (2019) introduce three variants to capture groupwise ranking errors. Kallus and Zhou (2019) propose a notion named xAUC, measuring the probability that positive samples from one group are ranked above negative samples from another, with the xAUC disparity quantifying cross-group differences.

To reduce xAUC disparities, many methods approximate the non-differentiable AUC objective using surrogate losses (Yao, Lin, and Yang 2023). Yang et al. (2023) frame the AUC objective as a zero-sum game problem so as to utilize a stochastic gradient optimization algorithm. Yao, Lin, and Yang (2023) approximate AUC objective and constraints with quadratic surrogate loss. However, such surrogate loss proxies may yield approximations that diverge significantly from the rank statistics they are intended to approximate (Rudin and Wang 2018). To avoid this issue, Cui et al. (2023) directly optimize bipartite rankings by greedily reordering cross-group instances to balance AUC and fairness, though at the cost of high computational complexity. Our FairPOT relates xAUC disparities with the divergence between score distributions of different groups, and applies optimal transport to reduce these disparities without retraining or complex reordering.

Partial AUC Optimization

While AUC has been widely used to measure the discriminative ability of a scoring function, in certain scenarios, partial AUC, which measures partial area under ROC curve, can be a more practically informative metric (Dodd and Pepe 2003; Carrington et al. 2020). Compared to AUC, partial AUC can be more appropriate for certain real-life scenarios (Narasimhan and Agarwal 2013a), such as biomedical screening (Zhu et al. 2024) in clinical practice and Top-K ranking (Shi et al. 2024) in recommendation systems.

Optimization of partial AUC has also been thoroughly explored in the literature. Some early works use indirect methods to relate objectives to partial area under the ROC curve, rather than directly optimizing partial AUC, including p-norm push (Rudin 2009) and infinite push (Agarwal 2011; Li, Jin, and Zhou 2014). Narasimhan and Agarwal (2013a,b) propose to use structural SVM to optimize the surrogate objective of partial AUC for linear models. Rudin and Wang (2018) employ mixed-integer programming to directly optimize partial AUC. Recent methods (Yang et al. 2021; Yao, Lin, and Yang 2022; Zhu et al. 2022) are based on stochastic gradient descent with approximate pairwise surrogate objectives which are amenable to deep learning. Despite growing

interest in partial AUC, existing work has focused almost exclusively on model performance. Our work fills an important gap by incorporating fairness into partial AUC optimization, aiming to balance cross-group fairness and predictive performance in top-risk regions.

Optimal Transport for Fairness

Optimal transport has been widely used in the literature to achieve fairness (Gordaliza et al. 2019; Jiang et al. 2020; Silvia et al. 2020; Zehlike, Hacker, and Wiedemann 2020; Laclau et al. 2021; Buyl and De Bie 2022; Xian, Yin, and Zhao 2023; Han et al. 2024; Ghassemi et al. 2025). Relevant methods span all three categories, i.e., pre-processing, in-processing, and post-processing. In pre-processing methods, optimal transport mainly aims to move features of different groups to their Wasserstein barycenter so as to minimize the deviation from original features (Gordaliza et al. 2019). In-processing methods incorporate fairness into the training process by adding a regularization term that uses the Wasserstein distance to penalize differences between the score distributions of different groups (Buyl and De Bie 2022). Post-processing methods typically focus on score distributions of different groups. For instance, Jiang et al. (2020) propose to conduct quantile matching based on the Wasserstein-1 distance across different groups. Our FairPOT method partially aligns the scores across groups, allowing for flexible, model-agnostic adjustments to balance fairness and accuracy.

Fairness-Accuracy Trade-off

Alongside efforts to achieve fairness, recent studies have also explored the trade-off between fairness and accuracy (Kleinberg, Mullainathan, and Raghavan 2016), motivating the need for mechanisms that balance both. A common approach leverages Pareto optimality (Mas-Colell et al. 1995) to characterize solutions where improving one objective necessarily compromises another (Xiao et al. 2017; Martinez, Bertran, and Sapiro 2020; Ge et al. 2022; Wei and Niethammer 2022; Xu and Strohmer 2023). Different studies interpret Pareto optimality from varying perspectives. Martinez, Bertran, and Sapiro (2020) interpret this as "no unnecessary harm" to any group. A more general formulation defines Pareto optimality in terms of the trade-off between fairness and accuracy (Xiao et al. 2017; Ge et al. 2022; Wei and Niethammer 2022; Xu and Strohmer 2023), requiring that no solution simultaneously worsens both metrics, as illustrated in Figure 2. While our method is not derived from these prior approaches, we adopt the Pareto optimality principle to characterize the trade-off frontier between fairness and accuracy. Specifically, we construct the Pareto frontier by filtering out all dominated points based on AUC and xAUC disparity, and evaluate our method by comparing it against the non-dominated points achieved by other fairness intervention methods.

Preliminaries

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $(X, Y, G) : \Omega \to \mathcal{X} \times \{0, 1\} \times \{a, b\}$ be a triplet of random variables

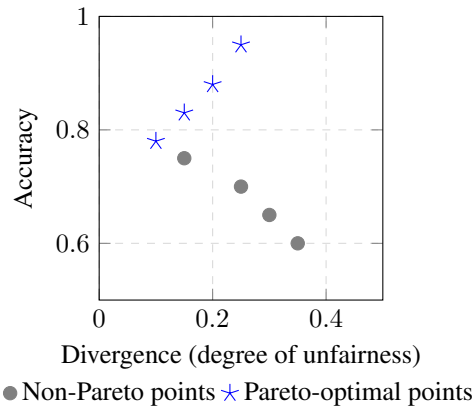

Figure 2: Illustration of the Pareto frontier: Pareto-optimal points (blue stars) dominate non-Pareto points (gray dots).

defined on this space. We denote by $\mathcal{D}$ the joint distribution of (X, Y, G) induced by $\mathbb{P}$, i.e., $(X, Y, G) \sim \mathcal{D}$. Here, $X \in \mathbb{R}^d$ is the feature vector, $Y \in \{0, 1\}$ is the binary label, and $G \in \{a, b\}$ is the sensitive attribute representing demographic groups where we want to impose fairness interventions. We focus on the case where the sensitive attribute G is binary, i.e., group a or b, which is common in many applications such as gender (e.g., male vs. female) or race (e.g., White vs. non-White). While sensitive attributes may naturally take on multiple categories, the binary setting allows for simpler group-wise comparisons and is widely adopted in the fairness literature. Denote a scoring function $f : \mathcal{X} \to [0, 1]$, where for each $x \in \mathcal{X}$, $f(x)$ is a real-valued score indicating the model belief of being labeled positively given x, i.e., $f(x) := \mathbb{P}[Y = 1 \mid X = x]$ (Jiang et al. 2020). We also denote $\mathbb{I}[\cdot]$ as the indicator function.

AUC and AUC-based Fairness

The AUC of a given score function f represents the probability that f assigns a strictly higher score to a randomly chosen positive instance than to a randomly chosen negative instance, which is

$$\mathrm{AUC}(f) \;=\; \mathbb{P}\big[f(X^+) > f(X^-)\big],$$

where X^+ and X^- are independently drawn from $\mathcal{D}$ conditioning on $Y = 1$ and $Y = 0$, respectively. Equivalently, we can rewrite the $\mathrm{AUC}(f)$ as follows

$$\mathrm{AUC}(f) \;=\; \mathbb{E}\Big[\mathbb{I}\big\{f(X^+) > f(X^-)\big\}\Big],$$

which calculates the fraction of positive-negative pairs that are correctly ranked by f. Consider now a finite sample dataset $\{(x_i, y_i, \hat{s}_i)\}_{i=1}^N$, where $\hat{s}_i = f(x_i)$, and denote the index sets

$$\mathcal{I}^+ = \{i \mid y_i = 1\}, \quad \mathcal{I}^- = \{j \mid y_j = 0\}.$$

Then the empirical AUC, $\widehat{\mathrm{AUC}}$, is defined by

$$\widehat{\mathrm{AUC}}(f) = \frac{\sum_{i \in \mathcal{I}^+} \sum_{j \in \mathcal{I}^-} \mathbb{I}\big[\hat{s}_i > \hat{s}_j\big]}{|\mathcal{I}^+| \;\cdot\; |\mathcal{I}^-|}.$$

We assume that $|\mathcal{I}^+| > 0$ and $|\mathcal{I}^-| > 0$; otherwise, $\widehat{\mathrm{AUC}}(f)$ is set to zero to indicate that no valid positive-negative comparisons can be made.

While AUC is widely used to evaluate the discriminative ability of a scoring function, it does not account for sensitive attributes or disparities between groups. In line with the metric proposed by Kallus and Zhou (2019), we use xAUC as the metric to measure how well f ranks positives in one group above negatives in another group. We refer to this quantity as the (cross-group) xAUC, highlighting that it compares samples across different demographic groups. Define

$$\mathrm{xAUC}_{a\to b}(f) \;=\; \mathbb{P}\big[f(X_a^+) > f(X_b^-)\big],$$

where X_a^+ and X_b^- are independently drawn from $\mathcal{D}$ conditioning on the events $\{Y=1, G=a\}$ and $\{Y=0, G=b\}$, respectively. Analogously, we define $\mathrm{xAUC}_{b\to a}(f)$ as the probability that a positive instance from group b receives a higher score than a negative instance from group a, i.e.,

$$\mathrm{xAUC}_{b\to a}(f) \;=\; \mathbb{P}\big[f(X_b^+) > f(X_a^-)\big].$$

To quantify the degree of fairness across groups, we define the disparity between xAUC as

$$\Delta\mathrm{xAUC}(a,b) = \big|\mathrm{xAUC}_{a\to b}(f) - \mathrm{xAUC}_{b\to a}(f)\big|.$$

Specifically, we say that xAUC-based fairness is achieved with a tolerance degree of $\varepsilon \geq 0$ if $\Delta\mathrm{xAUC}(a,b) \leq \varepsilon$ for a given $\varepsilon \geq 0$ (Yao, Lin, and Yang 2023). Empirically, let

$$\mathcal{I}_a^+ = \{\, i \mid y_i = 1, g_i = a\}, \quad \mathcal{I}_b^- = \{\, j \mid y_j = 0, g_j = b\}$$

denote the sets of positive samples from group a and negative samples from group b, respectively. Then the empirical xAUC is defined as

$$\widehat{\mathrm{xAUC}}_{a\to b}(f) = \frac{\sum_{i\in\mathcal{I}_a^+}\sum_{j\in\mathcal{I}_b^-}\mathbb{I}[\hat{s}_i > \hat{s}_j]}{|\mathcal{I}_a^+| \;\cdot\; |\mathcal{I}_b^-|}.$$

Hereby, we also assume that $|\mathcal{I}_a^+| > 0$ and $\big|\mathcal{I}_b^-\big| > 0$; otherwise, $\widehat{\mathrm{xAUC}}_{a\to b}(f)$ is set to zero to reflect the absence of valid pairwise comparisons. Similarly, $\widehat{\mathrm{xAUC}}_{b\to a}(f)$ is defined by swapping a and b. The empirical cross-group xAUC disparity is then given by

$$\Delta\widehat{\mathrm{xAUC}}(a,b) = \big|\widehat{\mathrm{xAUC}}_{a\to b}(f) - \widehat{\mathrm{xAUC}}_{b\to a}(f)\big|.$$

Partial AUC and Partial AUC-based Fairness

While a high AUC reflects good global separation of positive and negative instances, it does not necessarily imply strong ability to distinguish between classes within localized regions of the score distribution. A model may attain a high AUC by correctly ranking many instances in the middle range, while still failing to separate positives and negatives among the top-scoring individuals—who are often critical in high-stakes applications. For instance, in medical diagnosis, individuals with higher predicted risk scores require more attention and prioritization.

Partial AUC is a refinement of AUC that focuses on specific regions of the ROC curve that are practically more important for decision-making. Popular definitions of partial AUC fall into two categories: (i) restricting attention to a specified range of false positive rates (FPR) (Dodd and Pepe 2003; Narasimhan and Agarwal 2013a; Iwata, Fujino, and Ueda 2020; Zhu et al. 2022), and (ii) simultaneously restricting both FPR and true positive rates (TPR) (Yang et al. 2019, 2021; Chaibub Neto et al. 2024; Shi et al. 2024). For example, Dodd and Pepe (2003) define partial AUC as the area under the ROC curve restricted to $\mathrm{FPR} \leq \alpha$, under the rationale that in some domains (e.g., medical diagnosis), we cannot afford a high false alarm rate.

However, it is often unclear how to choose these ranges of FPR or TPR. For example, in a medical screening task where the disease rate is 3%, doctors may want to focus on the top 10% of patients with the highest risk scores—about three times the base rate. This top percentage is easier to decide in practice, while FPR or TPR values are harder to connect to real needs. Motivated by this, we propose a new notion of partial AUC defined over the top-α scoring region. Rather than integrating over FPR/TPR, we restrict attention to instances whose predicted scores fall within the top-α quantile of the score distribution.

Denote the α-quantile of the score distribution induced by f as t_α, i.e., $\mathbb{P}[f(X) \geq t_\alpha] = \alpha$ with X drawn from the marginal distribution of $\mathcal{D}$ over features. We consider independent random draws $X^+ \sim \mathcal{D}_1$ and $X^- \sim \mathcal{D}_0$, where $\mathcal{D}_1$ and $\mathcal{D}_0$ denote the conditional distributions of X given $Y=1$ and $Y=0$, respectively. Based on the joint sampling of (X^+, X^-), the partial AUC over the top-α region is defined as

$$\begin{aligned} &p\mathrm{AUC}(f;\alpha) \\ &= \mathbb{P}\Big[f(X^+) > f(X^-) \mid f(X^+) \geq t_\alpha, f(X^-) \geq t_\alpha\Big]. \end{aligned}$$

In other words, it measures how well the score function distinguishes positives from negatives among the top-scoring instances. When $\mathbb{P}[f(X^-) \geq t_\alpha] = 0$, we set pAUC to 1; when $\mathbb{P}[f(X^+) \geq t_\alpha] = 0$, we set it to 0, reflecting perfect or failed separation in the top-α region. Empirically, let the predicted scores be $\hat{\boldsymbol{s}} = (\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_N)$ sorted in descending order, and define $N_\alpha = \lceil \alpha N \rceil$. Let $\mathcal{I}_{N_\alpha}^+ = \{i \mid y_i = 1,\ i \leq N_\alpha\}$ and $\mathcal{I}_{N_\alpha}^- = \{j \mid y_j = 0,\ j \leq N_\alpha\}$, with $N_\alpha^+ = |\mathcal{I}_{N_\alpha}^+|$ and $N_\alpha^- = |\mathcal{I}_{N_\alpha}^-|$. Then, the empirical partial AUC is defined as

$$\widehat{p\mathrm{AUC}}(f;\alpha) = \frac{\sum_{i\in\mathcal{I}_{N_\alpha}^+}\sum_{j\in\mathcal{I}_{N_\alpha}^-}\mathbb{I}[\hat{s}_i > \hat{s}_j]}{N_\alpha^+ \cdot N_\alpha^-}.$$

Alternatively, it can be written as

$$\widehat{p\mathrm{AUC}}(f;\alpha) = \frac{\sum_{i\in\mathcal{I}^+}\sum_{j\in\mathcal{I}^-}\mathbb{I}[\hat{s}_i > \hat{s}_j]\,\mathbb{I}[\hat{s}_j \geq \hat{s}_{N_\alpha}]}{N_\alpha^+ \cdot N_\alpha^-},$$

where $\hat{s}_{N_\alpha}$ denotes the N_α-th largest score. We also assume that $N_\alpha^+ > 0$ and $N_\alpha^- > 0$. Similarly, pxAUC-based fairness is considered satisfied with tolerance $\varepsilon \geq 0$ if $\Delta p\mathrm{xAUC}(a,b) \leq \varepsilon$,

We now focus on fairness among the top-ranked instances. Let X_a^+ and X_b^- be drawn from $\{X \mid Y=1, G=a\}$ and $\{X \mid Y=0, G=b\}$, respectively, both restricted to

$f(\cdot) \geq t_\alpha$. Define

$$
\begin{aligned}
&p\mathrm{xAUC}_{a\to b}(f;\alpha) \\
&= \mathbb{P}\big[f(X_a^+) > f(X_b^-) \mid f(X_a^+) \geq t_\alpha,\, f(X_b^-) \geq t_\alpha\big], \\
&p\mathrm{xAUC}_{b\to a}(f;\alpha) \\
&= \mathbb{P}\big[f(X_b^+) > f(X_a^-) \mid f(X_b^+) \geq t_\alpha,\, f(X_a^-) \geq t_\alpha\big],
\end{aligned}
$$

and the partial fairness gap as

$$\Delta p\mathrm{xAUC}(a,b;\alpha) = |p\mathrm{xAUC}_{a\to b}(f;\alpha) - p\mathrm{xAUC}_{b\to a}(f;\alpha)|.$$

Empirically, for the top-N_α instances in each group, define index sets $\mathcal{I}^+_{N_\alpha,a} = \{i \mid y_i = 1, g_i = a, i \leq N_\alpha\}$ and $\mathcal{I}^-_{N_\alpha,b} = \{j \mid y_j = 0, g_j = b, j \leq N_\alpha\}$. Then, the empirical partial cross-group xAUCs are computed as

$$\widehat{p\mathrm{xAUC}}_{a\to b}(f;\alpha) = \frac{\sum_{i\in\mathcal{I}^+_{N_\alpha,a}} \sum_{j\in\mathcal{I}^-_{N_\alpha,b}} \mathbb{I}[\hat{s}_i > \hat{s}_j]}{|\mathcal{I}^+_{N_\alpha,a}| \cdot |\mathcal{I}^-_{N_\alpha,b}|},$$

with $\widehat{p\mathrm{xAUC}}_{b\to a}(f;\alpha)$ defined analogously. We also assume $|\mathcal{I}^+_{N_\alpha,a}| > 0$ and $|\mathcal{I}^-_{N_\alpha,b}| > 0$; if not, the corresponding empirical estimates are set to zero. The corresponding empirical partial xAUC disparity is

$$\Delta\widehat{p\mathrm{xAUC}}(a,b;\alpha) = |\widehat{p\mathrm{xAUC}}_{a\to b}(f;\alpha) - \widehat{p\mathrm{xAUC}}_{b\to a}(f;\alpha)|.$$

Optimal Transport

In this work, we focus on optimal transport in discrete settings (Monge 1781; Kantorovich 1942). Specifically, we consider two empirical measures

$$\mu = \sum_{i=1}^{n} u_i \delta_{z_i}, \quad \nu = \sum_{j=1}^{m} v_j \delta_{z'_j},$$

where δ_z denotes the Dirac delta measure at location z, $\{z_i\}_{i=1}^n$ and $\{z'_j\}_{j=1}^m$ are the support points, and u, v are their associated probability mass vectors (typically uniform). The objective is to find a coupling $\gamma \in \mathbb{R}^{n\times m}$ that transports μ to ν while minimizing the total transport cost. Formally, the discrete optimal transport problem solves

$$\gamma^* = \arg\min_{\gamma\in\Gamma(\mu,\nu)} \sum_{i=1}^{n}\sum_{j=1}^{m} \gamma_{i,j} C_{i,j}, \tag{1}$$

where $\Gamma(\mu,\nu) = \{\gamma \geq 0 \mid \gamma \mathbf{1}_m = u,\ \gamma^\top \mathbf{1}_n = v\}$ is the set of couplings with prescribed marginals, and $C_{i,j} = c(z_i, z'_j)$ defines the cost of transporting mass from z_i to z'_j. In our setting, we use the squared Euclidean distance as the cost, i.e., $C_{i,j} = (z_i - z'_j)^2$. Once the optimal coupling γ^* is found, we compute the barycentric projection (e.g., Seguy et al. (2017)) by

$$\tilde{z}_i = n \sum_{j=1}^{m} \gamma^*_{i,j} z'_j, \quad \forall i \in \{1,\ldots,n\}. \tag{2}$$

The barycentric projection defines a mapping from z_i toward z'_j based on the optimal coupling γ^*.

Fair Proportional Optimal Transport (FairPOT)

Existing studies focusing on AUC fairness (Fong et al. 2021; Yang et al. 2023; Yao, Lin, and Yang 2023; Cui et al. 2023) face a fundamental challenge: enforcing AUC fairness often leads to a deterioration in overall AUC performance. To address this challenge, we propose Fair Proportional Optimal Transport (FairPOT), a model agnostic post-processing method that employs optimal transport to align score distributions across different demographic groups to achieve fairness in terms of AUC and partial AUC with a tuning parameter $\lambda \in [0,1]$ to control the fairness-vs-accuracy trade-off.

Problem Definition

We study two settings where FairPOT can be applied: the standard (global) AUC setting and the partial AUC setting. The key difference lies in the scope of scores used for fairness intervention and evaluation as illustrated in Figure 3. In the global case, the entire score distribution is considered, whereas in the partial case, only the top-ranked (e.g., top-α quantile) region is evaluated and adjusted. In both cases, our goal is to trace the trade-off curve between fairness and accuracy by partially aligning scores of the disadvantaged group via optimal transport. Note that our fairness intervention is model-agnostic—it can be applied to any trained model, including black-box predictors. Our definition of the disadvantaged group and its relation to score distributions is further discussed in the Appendix.

Global AUC Case. Recall that we focus on binary sensitive attributes $G \in a, b$. Let $\hat{\boldsymbol{s}}_a = (\hat{s}_{a,1}, \ldots, \hat{s}_{a,n_a})$ and $\hat{\boldsymbol{s}}_b = (\hat{s}_{b,1}, \ldots, \hat{s}_{b,n_b})$ denote the predicted score vectors for groups a and b, respectively, obtained from scoring function f. Each element $\hat{s}_{a,i} = f(x_i) \in [0,1]$ represents the predicted risk score for the i-th instance in group a, and similarly for group b.

We use $\widehat{\mathrm{AUC}}(f)$ and $\Delta\widehat{\mathrm{xAUC}}(a,b)$ to measure predictive performance and fairness. Our objective is to construct a parametric family of partially transported scores $\{\tilde{\boldsymbol{s}}_b^\lambda\}_{\lambda\in[0,1]}$, where λ controls the degree of intervention to balance these two goals. Specifically, as λ varies, our goal is to obtain a trade-off curve in the $(\Delta\widehat{\mathrm{xAUC}}, \widehat{\mathrm{AUC}})$ space.

Partial AUC Case. In many real-world applications, such as screening or triage, decisions are based on only a small fraction of individuals with the highest predicted scores. To reflect this, we consider evaluating both performance and fairness within the top-α region of the score distribution.

Let $\alpha \in [0,1]$ denote the fraction of top-ranked instances used for evaluation. We compute the partial AUC $\widehat{p\mathrm{AUC}}(f;\alpha)$ and partial xAUC disparity $\Delta\widehat{p\mathrm{xAUC}}(a,b;\alpha)$ over this region. Similar to the global case, our objective is to construct a series of transformed score vectors $\tilde{\boldsymbol{s}}_b^{\lambda,\alpha}$ by applying optimal transport only within the top-α subset of scores. As λ varies, our goal is to produce a family of adjusted scores that define a trade-off curve in the $(\Delta\widehat{p\mathrm{xAUC}}, \widehat{p\mathrm{AUC}})$ space. This partial-view analysis re-

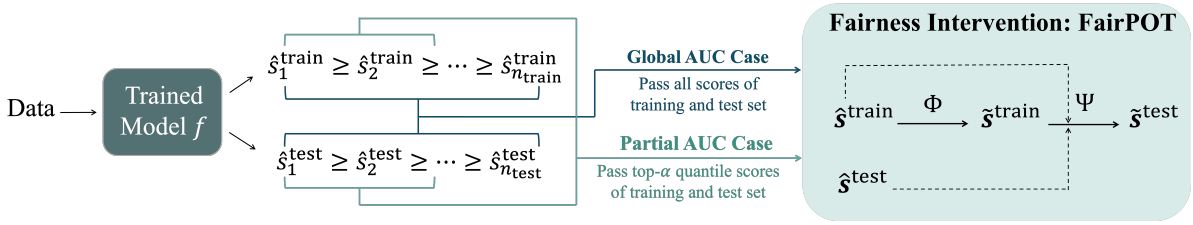

Figure 3: FairPOT framework. The partial transport Φ (Eq. (3)) maps training scores of group b to $\tilde{\boldsymbol{s}}_b^{\lambda,\text{train}}$, which are then used in the interpolation mapping Ψ (Eq. (4)) to adjust test scores.

flects realistic constraints where interventions are focused on high-risk populations.

Method

Global AUC Case. We apply optimal transport to align the distributions of scores of different groups. Without loss of generality, we denote a as the advantaged group and b as the disadvantaged group. We treat the predicted score vectors $\hat{\boldsymbol{s}}_a$ and $\hat{\boldsymbol{s}}_b$ as support points of the empirical score distributions. Specifically, we define the empirical marginals as $\mu_a = \frac{1}{n_a}\sum_{i=1}^{n_a}\delta_{\hat{s}_{a,i}}, \mu_b = \frac{1}{n_b}\sum_{j=1}^{n_b}\delta_{\hat{s}_{b,j}}$, where n_a, n_b are the number of samples in groups a and b, respectively. The optimal transport plan $\gamma^* \in \Gamma(\mu_b, \mu_a)$ is obtained by solving (1). Once the optimal plan γ^* is obtained, we compute the transported scores $\tilde{\boldsymbol{s}}_b$ by applying the barycentric projection defined in (2).

Applying the full transport map moves the scores of group b onto the support points of group a, such that each adjusted score of group b lies within the convex hull of scores from group a. This alignment substantially reduces the distributional gap between group a and group b, compared to the original unadjusted scores. Intuitively, by transporting each score in group b toward the support of group a, the two distributions become more closely aligned in shape and range. This reduces mismatches in score comparisons between groups, which helps improve symmetry in pairwise rankings. Symmetry ensures that positives from one group are not consistently ranked below negatives from another. However, altering the score distribution in this manner may distort the global ordering between positive and negative instances, leading to potential degradation in overall AUC. For example, a positive instance from the advantaged group may initially have a higher score than a negative instance from the disadvantaged group, but after transport, the adjusted score of the negative instance may exceed that of the positive one—resulting in incorrect ranking and reduced AUC. To flexibly balance the trade-off between fairness and accuracy, we introduce a parameter $\lambda \in [0,1]$ that controls the fraction of scores to be transported. Specifically, we sort the original scores $\hat{\boldsymbol{s}}_b$ in descending order and formally define the partial transport mapping Φ as

$$\Phi : (\hat{\boldsymbol{s}}_b; \gamma^*, \lambda) \longmapsto \tilde{\boldsymbol{s}}_b^{\lambda}, \tag{3}$$

Algorithm 1: FairPOT: Global AUC Case.

Input: Predicted scores $\hat{\boldsymbol{s}}^{\text{train}}$, $\hat{\boldsymbol{s}}^{\text{test}}$; group labels $\boldsymbol{g}^{\text{train}}$, $\boldsymbol{g}^{\text{test}}$; a set of discrete trade-off parameters $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_L\} \subseteq [0,1]$.
Output: Transported scores $\tilde{\boldsymbol{s}}^{\lambda,\text{train}}$, $\tilde{\boldsymbol{s}}^{\lambda,\text{test}}$ for each $\lambda \in \Lambda$; Pareto frontier.
Sort $(\hat{\boldsymbol{s}}^{\text{train}}, \boldsymbol{g}^{\text{train}})$ in descending order of $\hat{\boldsymbol{s}}^{\text{train}}$.
Sort $(\hat{\boldsymbol{s}}^{\text{test}}, \boldsymbol{g}^{\text{test}})$ in descending order of $\hat{\boldsymbol{s}}^{\text{test}}$.
Extract group-specific scores $\hat{\boldsymbol{s}}_a^{\text{train}} \leftarrow \{\hat{s}_i^{\text{train}} : g_i^{\text{train}} = a\}$ and $\hat{\boldsymbol{s}}_b^{\text{train}} \leftarrow \{\hat{s}_i^{\text{train}} : g_i^{\text{train}} = b\}$.
Define empirical distributions $\mu_a^{\text{train}} \leftarrow \frac{1}{n_a}\sum_{i=1}^{n_a}\delta_{\hat{s}_{a,i}^{\text{train}}}$ and $\mu_b^{\text{train}} \leftarrow \frac{1}{n_b}\sum_{j=1}^{n_b}\delta_{\hat{s}_{b,j}^{\text{train}}}$.
Define cost matrix $C_{i,j} \leftarrow (\hat{s}_{b,i}^{\text{train}} - \hat{s}_{a,j}^{\text{train}})^2$.
Compute optimal transport plan

$$\gamma^* \leftarrow \arg\min_{\gamma \in \Gamma(\mu_b^{\text{train}}, \mu_a^{\text{train}})} \sum_{i,j} \gamma_{i,j} C_{i,j}.$$

for each $\lambda \in \Lambda$ **do**
  Transport train scores $\tilde{\boldsymbol{s}}_b^{\lambda,\text{train}} \leftarrow \Phi(\hat{\boldsymbol{s}}_b^{\text{train}}; \gamma^*, \lambda)$.
  Map test scores $\tilde{\boldsymbol{s}}_b^{\lambda,\text{test}} \leftarrow \Psi(\hat{\boldsymbol{s}}_b^{\text{train}}; \tilde{\boldsymbol{s}}_b^{\lambda,\text{train}}, \hat{\boldsymbol{s}}_b^{\text{test}})$.
  Construct full transported scores

$$\tilde{\boldsymbol{s}}^{\lambda,\text{train}} \leftarrow \text{Merge}(\hat{\boldsymbol{s}}_a^{\text{train}}, \tilde{\boldsymbol{s}}_b^{\lambda,\text{train}}),$$

$$\tilde{\boldsymbol{s}}^{\lambda,\text{test}} \leftarrow \text{Merge}(\hat{\boldsymbol{s}}_a^{\text{test}}, \tilde{\boldsymbol{s}}_b^{\lambda,\text{test}}).$$

  Evaluate $\widehat{\text{AUC}}, \widehat{\Delta\text{xAUC}}$ on $\tilde{\boldsymbol{s}}^{\lambda,\text{test}}$.
end for
Identify non-dominated pairs of $(\widehat{\Delta\text{xAUC}}, \widehat{\text{AUC}})$ to form Pareto frontier across all $\lambda \in \Lambda$.

where

$$\tilde{s}_{b,i}^{\lambda} = \begin{cases} n_b \sum_{j=1}^{n_a} \gamma_{ji}^* \hat{s}_{a,j}, & \forall i \in \mathcal{I}_{b,\lambda}^{(\text{top})}, \\ \hat{s}_{b,i}, & \forall i \in \{1,\ldots,n_b\} \text{ and } i \notin \mathcal{I}_{b,\lambda}^{(\text{top})}, \end{cases}$$

and $n_{b,\lambda} = \lceil \lambda\, n_b \rceil$, $\mathcal{I}_{b,\lambda}^{(\text{top})} = \left\{ i : \hat{s}_{b,i} \leq \hat{s}_{b,n_{b,\lambda}} \right\}$.

As λ varies from 0 to 1, we obtain a family of post-processed score vectors $\tilde{\boldsymbol{s}}_b^{\lambda}$, interpolating between no transport ($\lambda = 0$) and full transport ($\lambda = 1$). For each $\lambda \in [0,1]$, we evaluate the corresponding empirical AUC and fairness metrics based on the transformed scores $\tilde{\boldsymbol{s}}_b^{\lambda}$. Specifically, we

compute $\big(\widehat{\mathrm{AUC}}(\tilde{\boldsymbol{s}}_b^{\lambda}),\ \widehat{\Delta\mathrm{xAUC}}(a,b;\tilde{\boldsymbol{s}}_b^{\lambda})\big)$, which traces a curve in the fairness–accuracy plane as λ varies. To formally characterize the optimal trade-offs, we adopt the principle of Pareto optimality. A point is Pareto-optimal if no other point achieves both higher AUC and lower fairness gap, with strict improvement in at least one. In practice, we identify the Pareto frontier by removing any point that is dominated by another, meaning there exists a point with better AUC and better fairness. The resulting curve captures the best achievable balance between AUC and fairness for different levels of transport.

Note that the post-processed scores are only defined for training data points. To generalize the transport to unseen testing data, we follow the strategy of Cui et al. (2023) by applying a piecewise linear interpolation between the original and transported training scores. Specifically, given original training scores $\widehat{\boldsymbol{s}}_b^{\text{train}}$ and transported counterparts $\tilde{\boldsymbol{s}}_b^{\lambda,\text{train}}$, we define a piecewise interpolation map Ψ that takes test scores $\widehat{\boldsymbol{s}}_b^{\text{test}}$ as input and outputs transported test scores $\tilde{\boldsymbol{s}}_b^{\lambda,\text{test}}$

$$\Psi : \left(\widehat{\boldsymbol{s}}_b^{\text{test}};\ \widehat{\boldsymbol{s}}_b^{\text{train}},\ \tilde{\boldsymbol{s}}_b^{\lambda,\text{train}}\right) \longmapsto \tilde{\boldsymbol{s}}_b^{\lambda,\text{test}}. \tag{4}$$

Specifically, for each test sample i from group b with score $\widehat{s}_{b,i}^{\text{test}}$, we locate the neighboring scores $(\widehat{s}_{b,i_1}^{\text{train}}, \tilde{s}_{b,i_1}^{\lambda,\text{train}})$ and $(\widehat{s}_{b,i_2}^{\text{train}}, \tilde{s}_{b,i_2}^{\lambda,\text{train}})$ from the training set such that $\widehat{s}_{b,i_1}^{\text{train}} \geq \widehat{s}_{b,i}^{\text{test}} \geq \widehat{s}_{b,i_2}^{\text{train}}$. We then apply linear interpolation

$$\tilde{s}_{b,i}^{\lambda,\text{test}} = \tilde{s}_{b,i_1}^{\lambda,\text{train}} + \left(\tilde{s}_{b,i_2}^{\lambda,\text{train}} - \tilde{s}_{b,i_1}^{\lambda,\text{train}}\right) \cdot \frac{\widehat{s}_{b,i}^{\text{test}} - \widehat{s}_{b,i_1}^{\text{train}}}{\widehat{s}_{b,i_2}^{\text{train}} - \widehat{s}_{b,i_1}^{\text{train}}}.$$

If $\widehat{s}_{b,i}^{\text{test}}$ falls outside the range of training scores, we extrapolate using the closest boundary value. The full algorithm is provided in Algorithm 1.

We choose to apply optimal transport to the top-λ portion of the score distribution, as opposed to other options, e.g., randomly selecting a λ fraction of samples. The intuition is: assuming group b is disadvantaged relative to group a, the cross-group AUC disparity $\Delta\mathrm{xAUC}(a,b)$ arises typically because the scoring function fails to adequately distinguish positive instances from group b and negative instances from group a. A natural approach to mitigating this issue is to increase the scores of likely-positive instances in group b, thereby improving their ranking relative to the negative samples in group a, and reducing the xAUC gap. However, since ground-truth labels are not accessible during model deployment, we cannot directly identify and adjust only the positive instances. Instead, we approximate this objective by targeting the instances in group b with the highest predicted scores, under the assumption that these represent the most confident positive predictions from the model. This top-λ region of the score distribution serves as a proxy for true positives, allowing us to intervene in a targeted and utility-preserving manner. In contrast, randomly selecting a λ fraction of instances would more likely include both positives and negatives. Modifying the scores of negative instances introduces unnecessary noise because such changes do not help reduce xAUC disparity but instead distort the score distribution without effectively addressing the root cause of xAUC disparity.

Partial AUC Case. Let $\hat{\boldsymbol{s}}_a^{\alpha}$ and $\hat{\boldsymbol{s}}_b^{\alpha}$ denote the subsets of top-α scores from f for groups a and b, respectively. To reduce the partial xAUC gap $\widehat{\Delta p\mathrm{xAUC}}(a,b;\alpha)$, we apply proportional optimal transport to align the distribution of $\hat{\boldsymbol{s}}_b^{\alpha}$ to that of $\hat{\boldsymbol{s}}_a^{\alpha}$ parametrized by λ. Owing to the selected partial AUC objective, we anticipate that the resulting trade-off curve between partial AUC and fairness can outperform the normal classifiers.

Specifically, we define empirical distributions μ_a^{α} and μ_b^{α} over $\hat{\boldsymbol{s}}_a^{\alpha}$ and $\hat{\boldsymbol{s}}_b^{\alpha}$ using uniform weights. We solve the discrete optimal transport problem between μ_b^{α} and μ_a^{α} with a cost matrix based on squared Euclidean distance, obtaining the optimal coupling γ^* using (1). Then, we apply a partial barycentric transport mapping as in (3)

$$\Phi_{\alpha} : (\hat{\boldsymbol{s}}_b^{\alpha}; \gamma^*, \lambda) \longmapsto \tilde{\boldsymbol{s}}_b^{\lambda,\alpha}, \tag{5}$$

where $\lambda \in [0,1]$ controls the extent of transport. By varying λ from 0 to 1, we obtain a family of partially transported score vectors $\{\tilde{\boldsymbol{s}}_b^{\lambda,\alpha}\}_{\lambda\in[0,1]}$. For each λ, we compute the corresponding $(\widehat{p\mathrm{AUC}}, \widehat{\Delta p\mathrm{xAUC}})$, which collectively trace out the fairness–accuracy curve. To characterize the optimal trade-offs, we also adopt the principle of Pareto optimality by removing any point that is dominated.

The post-processed scores for unseen testing data are handled similarly as described previously. The transport map Φ_{α} is learned based on the top-α scores from the training set, namely, $(\hat{\boldsymbol{s}}_a^{\alpha,\text{train}}, \hat{\boldsymbol{s}}_b^{\alpha,\text{train}})$. We then apply the learned transport to the test set using a separate linear interpolation procedure as in (4)

$$\Psi_{\alpha} : \left(\hat{\boldsymbol{s}}_b^{\alpha,\text{test}}; \hat{\boldsymbol{s}}_b^{\alpha,\text{train}}, \tilde{\boldsymbol{s}}_b^{\lambda,\alpha,\text{train}}\right) \longmapsto \tilde{\boldsymbol{s}}_b^{\lambda,\alpha,\text{test}}. \tag{6}$$

For each $\lambda \in [0,1]$, the evaluation of partial AUC and fairness gap is conducted based on the transported test scores $\tilde{\boldsymbol{s}}_b^{\lambda,\alpha,\text{test}}$, along with the unaltered scores $\hat{\boldsymbol{s}}_a^{\alpha,\text{test}}$ from group a. The full algorithm is provided in Appendix Algorithm 2.

Numerical Experiments

We empirically evaluate the performance of our proposed framework in balancing between 1) AUC-fairness and 2) partial AUC–fairness. Specifically, in both cases, we use FairPOT to adjust scores. The main difference lies in whether transport is applied to all the scores, as in AUC, or the top-α scores, as in partial AUC.

Datasets

We conduct empirical experiments across three datasets: (i) synthetic data with controllable group-level bias, (ii) public benchmark datasets commonly used in fairness studies, and (iii) a real-world clinical dataset for autism diagnosis.

Synthetic Data. We generate a synthetic dataset with $N = 3000$ samples and $D = 5$ continuous features, evenly split between two demographic groups a (advantaged) and b (disadvantaged). Features are sampled from group-specific Gaussian distributions with different means and variances: $\mathcal{N}(0.8, 1.0^2)$ for group a and $\mathcal{N}(0.1, 1.0^2)$ for group b. Each

group is assigned its own logistic scoring function with independently drawn coefficients, and intercepts are iteratively calibrated to yield positive rates of approximately 0.3 for group a and 0.1 for group b. Binary labels are then sampled from the resulting sigmoid-transformed scores.

Public Benchmark Datasets. `Bank` (Moro, Cortez, and Rita 2014) contains 16 personal and financial attributes of 45,211 individuals contacted during a marketing campaign, with the task of predicting whether the client will subscribe to a term deposit. We select *age* as the sensitive attribute and define individuals aged ≤ 25 as the *advantaged* group and those aged > 25 as the *disadvantaged* group. This choice is motivated by both task relevance and statistical disparity: younger individuals tend to have higher subscription rates (25.6%) than older individuals (11.4%), suggesting a relative advantage in this setting. Although older individuals may possess more financial stability, their lower response rate in this context justifies our designation for fairness evaluation. `COMPAS` contains 7,214 criminal recidivism records of 12 features predicting whether an individual will reoffend within two years (Bao et al. 2021). We choose sensitive attribute as *gender*, with *female* as disadvantaged group and *male* as advantaged group. Males make up about 80.7% of the dataset and have a higher recidivism rate (47.3%) compared to females (35.7%). Although females are less likely to reoffend, they are often underrepresented and may be treated unfairly in decision-making processes.

Autism Clinical Data. The clinical dataset contains 43,945 longitudinal clinical records collected between January 2015 and December 2023. Each patient record includes 21 demographic features (e.g., age, sex) and 229 diagnosis features derived from the Clinical Classifications Software (CCS) taxonomy, a standardized grouping of ICD diagnostic codes. The prediction target is a binary indicator for autism spectrum disorder (ASD) diagnosis. Due to the higher prevalence of ASD in males compared to females, there is a significant imbalance in the label distribution across gender, with the positive rate for males at 3.1% and for females at 1.0%. This naturally occurring divergence results in bias towards female. Therefore, we choose sensitive attribute as *gender*, with *female* as disadvantaged group and *male* as advantaged group.

Baselines

We evaluate the performance of FairPOT in the context of post-processing, where it is most comparable in terms of implementation and assumptions. Pre-processing and in-processing methods often require access to raw features, model gradients, or retraining, making it difficult to establish a fair and consistent comparison with our proposed FairPOT. We compare our method against several representative post-processing baselines. These methods are chosen because they are model-agnostic, require no retraining, and aim to improve score-based group fairness:

- **Unadjusted**: Standard XGBoost (Chen and Guestrin 2016) without any fairness intervention. We choose XGBoost as the base model due to its strong empirical performance and widespread adoption in tabular data tasks. It serves as the scoring function for all methods to ensure consistent comparison.
- **Post-logit** (Kallus and Zhou 2019): A group-specific score transformation method originally introduced in the xAUC paper. It applies a scaled sigmoid function to the scores of the disadvantaged group b: $\tilde{s}_b = \sigma(\alpha \cdot s_b + \beta)$, where $\sigma(\cdot)$ is the sigmoid function, β is fixed, and α is a scaling parameter tuned to reduce training disparity (e.g., xAUC gap). The scores of group a remain unchanged. We implement this method using the public code from Cui et al. (2023).
- **Wasserstein Fair Post-processing** (Jiang et al. 2020): A distribution-matching method that aligns score distributions across groups via quantile matching using the Wasserstein-1 distance. Each score is mapped to its group-wise quantile index, and then pushed forward to a shared barycenter quantile. This method represents a different use of optimal transport compared to FairPOT, and is included to contrast its performance.
- **xOrder** (Cui et al. 2023): A ranking-based post-processing method that reorders scores across groups by directly optimizing a value function balancing AUC and fairness. It supports fine-grained pairwise adjustments and serves as a strong baseline for score-level fairness.

Evaluation Protocols

Global AUC Case: (i) For the synthetic and public fairness datasets, we randomly split the data into training and test sets with a 80/20 ratio. For the clinical autism dataset, we follow task-specific cohort selection criteria to split the data. (ii) We first train a base classifier using **XGBoost** on the training set and obtain predicted risk scores for both the training and test sets. The predicted scores from the training data are then separated into advantaged and disadvantaged groups based on the sensitive attribute. (iii) For **xOrder** and **FairPOT**, we first apply the post-processing procedure to the training scores of the disadvantaged group only, using the advantaged group as a reference. The learned transformation is then generalized to the test scores of the disadvantaged group via interpolation based on quantile matching, as described in the method section. For **Post-Logit**, the learned adjustment function is directly applied to the test scores of the disadvantaged group, while the advantaged group scores remain unchanged. For **Wasserstein fair**, both the disadvantaged and advantaged group scores are jointly transformed at test time by computing a Wasserstein barycenter between their score distributions. (iv) Finally, for each method, we choose a set of discrete trade-off parameters $\Lambda = \{0.0, 0.1, 0.2, \ldots, 1.0\}$ to generate a family of adjusted score outputs. Then we compute a trade-off curve between accuracy and fairness, i.e., AUC vs. xAUC gap, enabling comprehensive comparison of different methods under different fairness-accuracy regimes. (v) For synthetic and public datasets, we repeat all experiments 20 times with different random seeds and report the mean to account for variability across data splits; for Autism dataset, due to task-specific cohort selection criteria, we use a fixed test set and resample 20 times with replacement from test set for boot-

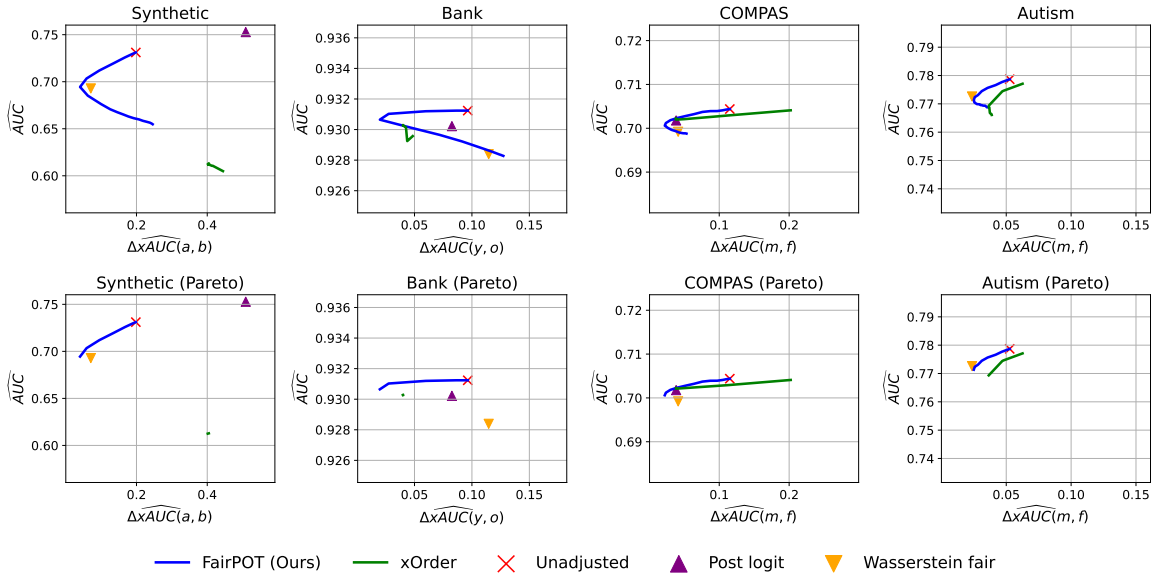

Figure 4: $\widehat{\text{AUC}} - \Delta\widehat{\text{xAUC}}$ trade-off.

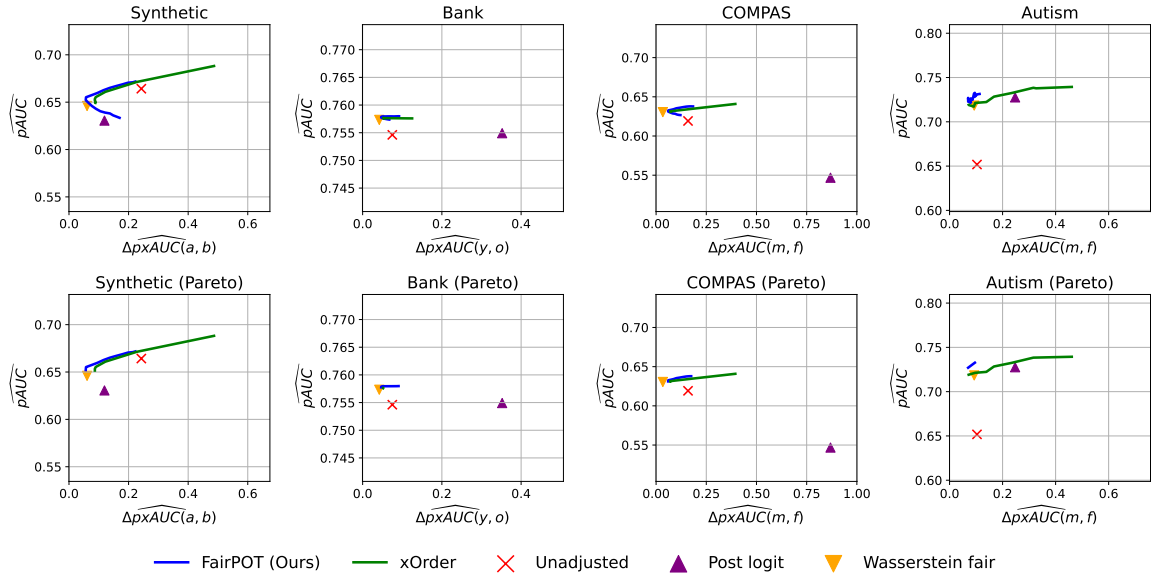

Figure 5: $\widehat{p\text{AUC}} - \Delta\widehat{p\text{xAUC}}$ trade-off. We set $\alpha = 0.3$ for Synthetic, Bank, and COMPAS datasets, and $\alpha = 0.1$ for Autism.

strap. We also report the results with standard error in the Appendix as in Figure 6.

Partial AUC Case: (i) We follow a similar data splitting strategy as in the first part, randomly dividing the dataset into training and test sets (80/20 split for synthetic and public datasets, task-driven split for clinical data). (ii) We train a base **XGBoost** on the training set to generate preliminary risk scores. These scores are then concatenated with the original features to form an augmented feature set. We use this augmented data to train a multilayer perceptron (MLP) model using a partial AUC-oriented objective as defined in Zhu et al. (2022); Yuan et al. (2023), e.g., a differentiable surrogate loss that encourages positive instances to rank above negatives within the top-α portion of the global ranking (Zhu et al. 2022; Yang 2022; Yuan et al. 2023). This yields a baseline model whose predictions already improve partial AUC. (iii) We then define the top-α region of interest based on the descending order of the optimized scores. Post-processing is performed only within this region. Specifically, we apply the proposed method and baselines to align the score distribution of the disadvantaged group in the top-α training region to that of the advantaged group, which is in

Table 1: Runtime (in seconds). Each entry shows mean ± std over 20 runs.

Dataset	Post logit	Wasserstein fair	xOrder	FairPOT (Ours)
Synthetic	0.195 ± 0.002	0.011 ± 0.000	3.687 ± 0.141	0.431 ± 0.026
Bank	0.445 ± 0.006	0.075 ± 0.005	276.700 ± 7.469	53.691 ± 1.975
COMPAS	0.230 ± 0.003	0.023 ± 0.001	20.314 ± 0.354	2.042 ± 0.049
Autism	1.242 ± 0.027	0.264 ± 0.015	2125.681 ± 18.857	360.578 ± 4.793

line with the procedures described in the global AUC case. (iv) For synthetic and public datasets, we repeat all experiments 20 times with different random seeds and report the mean to account for variability across data splits; for Autism dataset, due to task-specific cohort selection criteria, we use a fixed test set and resample 20 times with replacement from test set for bootstrap. We also report the results with standard error in the Appendix as in Figure 7.

Results

Improved Pareto Frontier for $\widehat{\mathrm{AUC}} - \Delta\widehat{\mathrm{xAUC}}$ Trade-off. As shown in Figure 4, FairPOT consistently achieves a superior or comparable Pareto frontier compared to baselines including xOrder, Post-logit adjustment, and Wasserstein fair across all datasets. In the Synthetic, the Bank, and the COMPAS datasets, the FairPOT curve strictly dominates other methods. For every level of fairness gap, FairPOT attains a higher AUC. In the Autism dataset, FairPOT dominates the baselines in most regions. Although Wasserstein fair achieves a lower fairness gap on Autism, it causes a substantial drop in AUC and lacks the flexibility to explore trade-offs between fairness and accuracy. Notably, the unadjusted point always lies strictly inside the FairPOT frontier, confirming that FairPOT can recover the original model by selecting $\lambda = 0$. Among all baselines, xOrder is the only one demonstrating some degree of fairness–accuracy trade-off. However, FairPOT consistently yields a better Pareto frontier than xOrder across all datasets. Moreover, we observe that the Pareto curve produced by xOrder does not always pass through the original unadjusted model performance point (i.e., the $(\widehat{\mathrm{AUC}}, \Delta\widehat{\mathrm{xAUC}})$ value before any post-processing). This suggests that xOrder is unable to fully recover the original scoring function by adjusting its tuning parameter, and may introduce greater deviations from the original model predictions compared to FairPOT. Also, xOrder is more sensitive to its tuning parameter, making it difficult to select an appropriate configuration in practice.

Simultaneous Improvements in $\widehat{p\mathrm{AUC}}$ and $\Delta\widehat{p\mathrm{xAUC}}$. As shown in Figure 5, FairPOT in the partial AUC setting demonstrates clear regions where both partial AUC and fairness are simultaneously improved compared to the unadjusted XGBoost. In the Bank, the COMPAS, and the Autism datasets, FairPOT consistently achieves joint improvements across almost the entire curve, highlighting the potential to move beyond the traditional trade-off between partial AUC and fairness in top-α scenarios. On Synthetic, FairPOT also identifies regions with simultaneous gains, although the improvements are more localized. While xOrder sometimes achieves slightly higher partial AUC values compared to FairPOT, these gains often come at the expense of fairness, with fairness gaps even exceeding those of the original (unadjusted) model. Such regions are less desirable in practice, as our goal is to improve fairness without sacrificing it relative to the starting point. In contrast, FairPOT provides more balanced improvements, ensuring that fairness is at least maintained or enhanced while promoting better partial AUC among top-scoring individuals.

Computation Efficiency. Experiments on Synthetic, Bank, and COMPAS datasets were conducted on a personal device (Apple M4 Pro, 24 GB RAM). Experiments on the Autism dataset were conducted on a secure remote server (Intel Xeon E5-2699 v4, 16 GB RAM), due to data access restrictions preventing local processing. As summarized in Table 1, FairPOT offers a practical balance between fairness improvement and computational efficiency. Compared to xOrder, FairPOT is significantly faster while achieving comparable or better fairness–accuracy trade-offs. While Post-logit adjustment and Wasserstein fair run faster, they fail to achieve good Pareto trade-offs across datasets. Thus, FairPOT strikes a favorable trade-off between runtime and effectiveness, making it a practical choice for real-world applications requiring both fairness and utility.

Towards Fair and Optimal Decision Making. Our proposed framework provides actionable insights for practical decision-making. For instance, on the Autism dataset, FairPOT in the partial AUC setting not only improves partial AUC but also reduces partial fairness disparity compared to the original XGBoost, enabling more equitable identification of high-risk individuals. In domains such as healthcare, screening, and criminal justice, where resource allocation is constrained, such improvements enable more effective and fair downstream interventions. We also conduct a more detailed analysis of how the chosen α can affect the decision-making for Autism data, as shown in the Appendix.

Conclusion

We propose *Fair Proportional Optimal Transport* (FairPOT), a simple and flexible post-processing method to improve fairness of risk scores while preserving model performance. By adjusting only part of the scores, FairPOT allows a clear trade-off between fairness and AUC. We also extend this method to partial AUC, focusing on top-risk regions important in real-world scenarios. Experiments on multiple datasets show that FairPOT performs better than existing post-processing methods, offering a good balance between fairness and utility with low computational cost.

Acknowledgements

This work is supported in part by The Duke Endowment (TDE) under grant #7262-SP and by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) under grant P50 HD093074.

References

Agarwal, S. 2011. The infinite push: A new support vector ranking algorithm that directly optimizes accuracy at the absolute top of the list. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, 839–850. SIAM.

Bao, M.; Zhou, A.; Zottola, S.; Brubach, B.; Desmarais, S.; Horowitz, A.; Lum, K.; and Venkatasubramanian, S. 2021. It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. *arXiv preprint arXiv:2106.05498*.

Borkan, D.; Dixon, L.; Sorensen, J.; Thain, N.; and Vasserman, L. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, 491–500.

Buyl, M.; and De Bie, T. 2022. Optimal transport of classifiers to fairness. *Advances in Neural Information Processing Systems*, 35: 33728–33740.

Carrington, A. M.; Fieguth, P. W.; Qazi, H.; Holzinger, A.; Chen, H. H.; Mayr, F.; and Manuel, D. G. 2020. A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. *BMC medical informatics and decision making*, 20: 1–12.

Caton, S.; and Haas, C. 2024. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.*, 56(7).

Chaibub Neto, E.; Yadav, V.; Sieberts, S. K.; and Omberg, L. 2024. A novel estimator for the two-way partial AUC. *BMC Medical Informatics and Decision Making*, 24(1): 57.

Chen, T.; and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Cui, S.; Pan, W.; Zhang, C.; and Wang, F. 2023. Bipartite ranking fairness through a model agnostic ordering adjustment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 13235–13249.

Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67–73.

Dodd, L. E.; and Pepe, M. S. 2003. Partial AUC estimation and regression. *Biometrics*, 59(3): 614–623.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Ensign, D.; Friedler, S. A.; Neville, S.; Scheidegger, C.; and Venkatasubramanian, S. 2018. Runaway feedback loops in predictive policing. In *Conference on fairness, accountability and transparency*, 160–171. PMLR.

Fong, H.; Kumar, V.; Mehrotra, A.; and Vishnoi, N. K. 2021. Fairness for auc via feature augmentation. *arXiv preprint arXiv:2111.12823*.

Ge, Y.; Zhao, X.; Yu, L.; Paul, S.; Hu, D.; Hsieh, C.-C.; and Zhang, Y. 2022. Toward pareto efficient fairness-utility trade-off in recommendation through reinforcement learning. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, 316–324.

Ghassemi, M.; Mishler, A.; Dalmasso, N.; Zhang, L.; Potluru, V. K.; Balch, T.; and Veloso, M. 2025. Auditing and Enforcing Conditional Fairness via Optimal Transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 16808–16816.

Gordaliza, P.; Del Barrio, E.; Fabrice, G.; and Loubes, J.-M. 2019. Obtaining fairness using optimal transport theory. In *International conference on machine learning*, 2357–2365. PMLR.

Han, Z.; Chen, C.; Zheng, X.; Li, M.; Liu, W.; Yao, B.; Li, Y.; and Yin, J. 2024. Intra-and inter-group optimal transport for user-oriented fairness in recommender systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8463–8471.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Iwata, T.; Fujino, A.; and Ueda, N. 2020. Semi-supervised learning for maximizing the partial AUC. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 4239–4246.

Jiang, R.; Pacchiano, A.; Stepleton, T.; Jiang, H.; and Chiappa, S. 2020. Wasserstein fair classification. In *Uncertainty in artificial intelligence*, 862–872. PMLR.

Kallus, N.; and Zhou, A. 2019. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. *Advances in neural information processing systems*, 32.

Kantorovich, L. 1942. On the transfer of masses (in Russian). In *Doklady Akademii Nauk*, volume 37, 227.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Laclau, C.; Redko, I.; Choudhary, M.; and Largeron, C. 2021. All of the fairness for edge prediction with optimal transport. In *International Conference on Artificial Intelligence and Statistics*, 1774–1782. PMLR.

Li, N.; Jin, R.; and Zhou, Z.-H. 2014. Top rank optimization in linear time. *Advances in neural information processing systems*, 27.

Martinez, N.; Bertran, M.; and Sapiro, G. 2020. Minimax pareto fairness: A multi objective perspective. In *International conference on machine learning*, 6755–6764. PMLR.

Mas-Colell, A.; Whinston, M. D.; Green, J. R.; et al. 1995. *Microeconomic theory*, volume 1. Oxford university press New York.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Monge, G. 1781. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, 666–704.

Moro, S.; Cortez, P.; and Rita, P. 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62: 22–31.

Narasimhan, H.; and Agarwal, S. 2013a. A structural SVM based approach for optimizing partial AUC. In *International Conference on Machine Learning*, 516–524. PMLR.

Narasimhan, H.; and Agarwal, S. 2013b. SVMpAUCtight: a new support vector method for optimizing partial AUC based on a tight convex upper bound. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 167–175.

Pessach, D.; and Shmueli, E. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3): 1–44.

Rajkomar, A.; Hardt, M.; Howell, M. D.; Corrado, G.; and Chin, M. H. 2018. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12): 866–872.

Rudin, C. 2009. The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list.

Rudin, C.; and Wang, Y. 2018. Direct learning to rank and rerank. In *International Conference on Artificial Intelligence and Statistics*, 775–783. PMLR.

Seguy, V.; Damodaran, B. B.; Flamary, R.; Courty, N.; Rolet, A.; and Blondel, M. 2017. Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*.

Shi, W.; Wang, C.; Feng, F.; Zhang, Y.; Wang, W.; Wu, J.; and He, X. 2024. Lower-left partial auc: An effective and efficient optimization metric for recommendation. In *Proceedings of the ACM Web Conference 2024*, 3253–3264.

Silvia, C.; Ray, J.; Tom, S.; Aldo, P.; Heinrich, J.; and John, A. 2020. A general approach to fairness with optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 3633–3640.

Suresh, H.; and Guttag, J. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9.

Wei, S.; and Niethammer, M. 2022. The fairness-accuracy Pareto front. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(3): 287–302.

Xian, R.; Yin, L.; and Zhao, H. 2023. Fair and optimal classification via post-processing. In *International conference on machine learning*, 37977–38012. PMLR.

Xiao, L.; Min, Z.; Yongfeng, Z.; Zhaoquan, G.; Yiqun, L.; and Shaoping, M. 2017. Fairness-aware group recommendation with pareto-efficiency. In *Proceedings of the eleventh ACM conference on recommender systems*, 107–115.

Xu, S.; and Strohmer, T. 2023. Fair data representation for machine learning at the Pareto frontier. *Journal of Machine Learning Research*, 24(331): 1–63.

Yang, H.; Lu, K.; Lyu, X.; and Hu, F. 2019. Two-way partial AUC and its properties. *Statistical methods in medical research*, 28(1): 184–195.

Yang, T. 2022. Algorithmic Foundations of Empirical X-Risk Minimization. *arXiv preprint arXiv:2206.00439*.

Yang, Z.; Ko, Y. L.; Varshney, K. R.; and Ying, Y. 2023. Minimax auc fairness: Efficient algorithm with provable convergence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 11909–11917.

Yang, Z.; Xu, Q.; Bao, S.; He, Y.; Cao, X.; and Huang, Q. 2021. When all we need is a piece of the pie: A generic framework for optimizing two-way partial AUC. In *International Conference on Machine Learning*, 11820–11829. PMLR.

Yao, Y.; Lin, Q.; and Yang, T. 2022. Large-scale optimization of partial auc in a range of false positive rates. *Advances in Neural Information Processing Systems*, 35: 31239–31253.

Yao, Y.; Lin, Q.; and Yang, T. 2023. Stochastic methods for auc optimization subject to auc-based fairness constraints. In *International Conference on Artificial Intelligence and Statistics*, 10324–10342. PMLR.

Yuan, Z.; Zhu, D.; Qiu, Z.-H.; Li, G.; Wang, X.; and Yang, T. 2023. LibAUC: A Deep Learning Library for X-Risk Optimization. In *29th SIGKDD Conference on Knowledge Discovery and Data Mining*.

Zehlike, M.; Hacker, P.; and Wiedemann, E. 2020. Matching code and law: achieving algorithmic fairness with optimal transport. *Data Mining and Knowledge Discovery*, 34(1): 163–200.

Zhu, D.; Li, G.; Wang, B.; Wu, X.; and Yang, T. 2022. When AUC meets DRO: Optimizing partial AUC for deep learning with non-convex convergence guarantee. In *International Conference on Machine Learning*, 27548–27573. PMLR.

Zhu, X.; Ren, X.; Shi, W.; Wang, C.; Liu, X.; Liu, Y.; Tao, T.; and Feng, F. 2024. Improving Prostate Cancer Risk Prediction through Partial AUC Optimization. In *Companion Proceedings of the ACM Web Conference 2024*, 1170–1173.

Appendix

Appendix A. Group and Score Interpretation

In this work, we define the disadvantaged group not based on score magnitude or social identity alone, but rather in terms of model performance. Specifically, the group for which the model demonstrates worse discriminative ability in ranking positive instances above negative instances from the other group. Formally, this corresponds to lower xAUC, where the probability that a randomly chosen positive from the disadvantaged group ranks above a randomly chosen negative from the advantaged group is lower than the reverse.

This framing does not assume whether higher or lower scores are inherently advantageous. In some tasks, the disadvantaged group may exhibit higher average scores. For example, older individuals in financial risk prediction may be over-scored due to age-related biases, yet the model may still fail to correctly distinguish between true positives and negatives within this group. In other cases, the disadvantaged group may have lower average scores, as observed in recidivism prediction, where female individuals tend to receive lower risk scores, but the model may underperform in identifying true positives. FairPOT is designed to flexibly handle both types of scenarios. Since it focuses on improving relative ranking fairness through cross-group pairwise comparisons, it does not rely on assumptions about the absolute direction of scores.

Appendix B. Algorithm for Partial AUC Case of FairPOT

Algorithm 2 outlines the pseudocode for applying FairPOT in the partial AUC setting, where the fairness intervention is restricted to the top-scoring region of the output.

Appendix C. Implementation Details

Global AUC Case. In the global AUC setting, we use XGBoost (Chen and Guestrin 2016) as the base classifier, due to its consistently strong empirical performance and widespread adoption in structured/tabular data tasks as shown in Table 2. While we adopt XGBoost for consistency, it is important to note that FairPOT is a model-agnostic post-processing method. This means it can be applied to predictions from any model, including black-box models, without requiring retraining or model access.

Partial AUC Case. In the partial AUC setting, we focus on improving model performance and fairness within the top-ranked region of the score distribution. To this end, we adopt the pairwise DRO-style objective introduced by Zhu et al. (2022), which efficiently approximates partial AUC. The optimization problem is defined as

$$\min_{\mathbf{w}} \frac{1}{n_+} \sum_{\mathbf{x}_i \in \mathcal{S}_+} \lambda \log \left(\frac{1}{n_-} \sum_{\mathbf{x}_j \in \mathcal{S}_-} \exp\left(\frac{L(\mathbf{w}; \mathbf{x}_i, \mathbf{x}_j)}{\lambda} \right) \right)$$

Here, $\mathcal{S}_+$ and $\mathcal{S}_-$ represent the sets of positive and negative instances, and $L(\mathbf{w}; \mathbf{x}_i, \mathbf{x}_j)$ is a pairwise surrogate loss such as squared hinge loss. This objective prioritizes hard positive-negative pairs and concentrates learning in the top-risk region.

Additionally, we improve optimization by incorporating predicted scores from XGBoost as an extra feature input. Table 3 compares the performance of:

(a) unadjusted XGBoost scores,

(b) partial AUC optimization using only original features, and

(c) partial AUC optimization using both original features and XGBoost scores.

We apply fairness intervention on both global and partial AUC settings. However, we only optimize the model toward partial AUC, and not global AUC. This is because partial AUC optimization is more practical: in the partial setting, we can explicitly select instances from the top-α region to guide learning. In contrast, global AUC optimization requires reordering all instance pairs across the entire score distribution, which is algorithmically more challenging.

Algorithm 2: FairPOT: Partial AUC Case.

Input: Predicted scores $\hat{\boldsymbol{s}}^{\text{train}}$, $\hat{\boldsymbol{s}}^{\text{test}}$; group labels $\boldsymbol{g}^{\text{train}}$, $\boldsymbol{g}^{\text{test}}$; top region parameter α, a set of discrete trade-off parameters $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_L\} \subseteq [0, 1]$.

Output: Transported scores of top α region $\tilde{\boldsymbol{s}}^{\lambda,\alpha,\text{train}}$, $\tilde{\boldsymbol{s}}^{\lambda,\alpha,\text{test}}$ for each $\lambda \in \Lambda$; Pareto frontier.

Sort $\hat{\boldsymbol{s}}^{\text{train}}$ in descending order and take top N_α^{train} to obtain $\hat{\boldsymbol{s}}_b^{\alpha,\text{train}}$.

Sort $\hat{\boldsymbol{s}}_b^{\text{test}}$ in descending order and take top N_α^{test} to obtain $\hat{\boldsymbol{s}}_b^{\alpha,\text{test}}$.

Extract group-specific top scores $\hat{\boldsymbol{s}}_a^{\alpha,\text{train}}$, $\hat{\boldsymbol{s}}_b^{\alpha,\text{train}}$.

Define empirical distributions $\mu_a^{\alpha,\text{train}}$, $\mu_b^{\alpha,\text{train}}$ over top-α samples.

Solve optimal transport plan

$$\gamma^* \leftarrow \arg \min_{\gamma \in \Gamma(\mu_b^{\alpha,\text{train}}, \mu_a^{\alpha,\text{train}})} \sum_{i,j} \gamma_{i,j} (\hat{s}_{b,i}^{\alpha,\text{train}} - \hat{s}_{a,j}^{\alpha,\text{train}})^2.$$

for each $\lambda \in \Lambda$ **do**

Transport train scores

$$\tilde{\boldsymbol{s}}_b^{\lambda,\alpha,\text{train}} \leftarrow \Phi_\alpha(\hat{\boldsymbol{s}}_b^{\alpha,\text{train}}; \gamma^*, \lambda).$$

Map test scores

$$\Psi_\alpha : \left(\hat{\boldsymbol{s}}_b^{\alpha,\text{test}}; \hat{\boldsymbol{s}}_b^{\alpha,\text{train}}, \tilde{\boldsymbol{s}}_b^{\lambda,\alpha,\text{train}} \right) \longmapsto \tilde{\boldsymbol{s}}_b^{\lambda,\alpha,\text{test}}.$$

Construct full transported scores

$$\tilde{\boldsymbol{s}}^{\lambda,\alpha,\text{train}} \leftarrow \text{Merge}(\hat{\boldsymbol{s}}_a^{\alpha,\text{train}}, \tilde{\boldsymbol{s}}_b^{\lambda,\alpha,\text{train}}),$$

$$\tilde{\boldsymbol{s}}^{\lambda,\alpha,\text{test}} \leftarrow \text{Merge}(\hat{\boldsymbol{s}}_a^{\alpha,\text{test}}, \tilde{\boldsymbol{s}}_b^{\lambda,\alpha,\text{test}}).$$

Evaluate $\widehat{p\text{AUC}}$ and $\widehat{\Delta p\text{xAUC}}$ on $\tilde{\boldsymbol{s}}^{\lambda,\alpha,\text{test}}$.

end for

Identify Pareto frontier among evaluated $(\widehat{\Delta p\text{xAUC}}, \widehat{p\text{AUC}})$ points.

Table 2: AUC for different classifiers. Each entry shows mean ± std over 20 runs.

Classifier	Synthetic	Bank	COMPAS	Autism	Average
Logistic Regression	0.7254 ± 0.0202	0.9059 ± 0.0036	0.7207 ± 0.0134	0.7161 ± 0.0179	0.7670
Random Forest	0.7583 ± 0.0132	0.9281 ± 0.0025	0.6587 ± 0.0117	0.6346 ± 0.0189	0.7449
Gradient Boosting	$\mathbf{0.7670} \pm 0.0143$	0.9248 ± 0.0027	$\mathbf{0.7259} \pm 0.0118$	0.6714 ± 0.0155	0.7723
SVM	0.7363 ± 0.0136	0.9091 ± 0.0038	0.7223 ± 0.0134	0.6648 ± 0.0164	0.7581
Naive Bayes	0.7550 ± 0.0173	0.8083 ± 0.0062	0.6950 ± 0.0131	0.4805 ± 0.0084	0.6847
Decision Tree	0.5960 ± 0.0191	0.7019 ± 0.0079	0.6084 ± 0.0130	0.5253 ± 0.0102	0.6079
KNN	0.7129 ± 0.0108	0.8116 ± 0.0063	0.6608 ± 0.0125	0.5393 ± 0.0152	0.6812
Extra Trees	0.7577 ± 0.0131	0.9142 ± 0.0027	0.6423 ± 0.0112	0.7113 ± 0.0270	0.7564
XGBoost	0.7311 ± 0.0139	$\mathbf{0.9312} \pm 0.0024$	0.7044 ± 0.0105	$\mathbf{0.7731} \pm 0.0143$	**0.7850**

Table 3: pAUC evaluated from different method.

Method	Synthetic	Bank	COMPAS	Autism
(a)	0.6642	0.7546	0.6171	0.633
(b)	**0.6798**	0.7544	0.6153	0.573
(c)	0.6740	**0.7584**	**0.6383**	**0.749**

Appendix D. Extended Result Analysis

To further demonstrate the model-agnostic nature of FairPOT, we apply it to additional classifiers—Gradient Boosting (Figures 8, 9) and Random Forest (Figures 10, 11). For Gradient Boosting, FairPOT achieves Pareto-dominant or comparable trade-offs across datasets. On Synthetic and COMPAS, FairPOT consistently outperforms Wasserstein Fair and Post-logit, and matches or surpasses xOrder. In the Bank dataset, Post-logit shows lower fairness disparity but with a notable drop in AUC and no ability to trade off. On Autism, xOrder fails to improve fairness, while FairPOT retains flexibility to balance fairness and accuracy. For Random Forest, FairPOT again shows strong performance. On Synthetic, it clearly outperforms all baselines, while Post-logit increases unfairness. On Bank, FairPOT dominates. On COMPAS, it achieves both improved fairness and broader control over the fairness–accuracy balance. In the Autism dataset, where the unadjusted model already yields near-parity ($\Delta\mathrm{xAUC} \approx 0.01$), most baselines degrade fairness, whereas FairPOT is able to recover the original outputs with appropriately chosen λ. Overall, these results confirm the effectiveness and stability of FairPOT across different training models, supporting its practical value as a flexible and efficient post-processing solution.

Figure 6 and Figure 7 present the mean and standard error corresponding to the results shown in Figure 4 and Figure 5, respectively.

Figure 12 illustrates how $\widehat{p\mathrm{AUC}}$ and $\widehat{p\mathrm{xAUC}}$ vary with α, both with and without applying partial AUC optimization as described in the partial AUC case of the evaluation protocols. From Figure 12, we observe that under the original XGBoost model, both $\widehat{p\mathrm{AUC}}$ and $\widehat{p\mathrm{xAUC}}$ tend to increase as α grows. This indicates that the model performs relatively worse among higher-risk individuals (i.e., those with higher predicted scores), which is undesirable in practical settings, especially for the Autism dataset where the positive rate is only around 2%. After applying partial AUC optimization, however, this trend is reversed: both $\widehat{p\mathrm{AUC}}$ and $\widehat{p\mathrm{xAUC}}$ decrease as α increases. This suggests that the ability of the model to distinguish high-risk individuals has been effectively enhanced, precisely in the regions that matter most for clinical decision-making. Such improvement is particularly important in low-prevalence conditions like autism, where better score calibration in the high-risk region provides stronger support for thresholding decisions and offers greater flexibility for achieving fairness in downstream interventions.

Note that in this experiment, we apply the partial AUC optimization procedure as described in the partial AUC evaluation protocol. Therefore, the model used here differs from the base model used in the main experiments (which directly employ XGBoost without partial AUC tuning). This explains why the results for $\alpha = 1$ in Figure 12 differ from the global AUC results shown earlier. Although $\alpha = 1$ formally recovers the global AUC region, the optimization objective and resulting model are different, which leads to the observed discrepancy.

This highlights an important distinction: while the global AUC case evaluates the overall discrimination ability of base model, the partial AUC case (even with $\alpha = 1$) focuses on optimizing a specific risk region. Understanding this distinction helps unify the interpretation of FairPOT under different evaluation settings.

Appendix E. Clarifying Suboptimal λ Configurations in Trade-off Curves

As shown in Figures 4 and 5, we visualize the full performance curves of FairPOT across different values of $\lambda \in [0, 1]$ to illustrate the trade-off between fairness and accuracy. In this process, certain λ values may yield both lower $\widehat{\mathrm{AUC}}$ and higher $\Delta\widehat{\mathrm{xAUC}}$ (or their partial counterparts) compared to the unadjusted model ($\lambda = 0$). While these points may appear undesirable, we emphasize that they are included purely for completeness and transparency of the entire trade-off curve.

In practice, we do not select such suboptimal configura-

tions. As discussed in Section 4, FairPOT consistently yields a superior or comparable Pareto frontier compared to baseline methods, and always includes the original model as a special case with $\lambda = 0$. To accommodate diverse application needs, we present the full Pareto-optimal region, which represents the set of non-dominated trade-offs between fairness and accuracy. The final selection of a specific λ value should be made based on the decision maker's preference over fairness and utility, depending on the context of deployment.

Appendix F. Ablation: Switching Advantaged Group and Disadvantaged Group

Our main method applies FairPOT to shift the scores of the disadvantaged group b toward the support of the advantaged group a, followed by interpolation to test data. A natural question arises: does the direction of transport matter? That is, what if we switch group a and group b?

To answer this, we repeat our method on all four datasets (Synthetic, Bank, COMPAS, and Autism), reversing the transport direction. Specifically, we shift the training scores of group a toward the support of group b, while keeping the test-time interpolation procedure unchanged. From Figure 13 and Figure 14, we observe negligible differences between the two transport directions in terms of the $\widehat{\mathrm{AUC}}$–$\Delta\mathrm{x}\widehat{\mathrm{AUC}}$ trade-off, which suggests that either direction yields comparable fairness–accuracy outcomes. We adopt the convention of shifting the disadvantaged group in our main method to align with fairness literature, which emphasizes mitigating harm for disadvantaged populations.

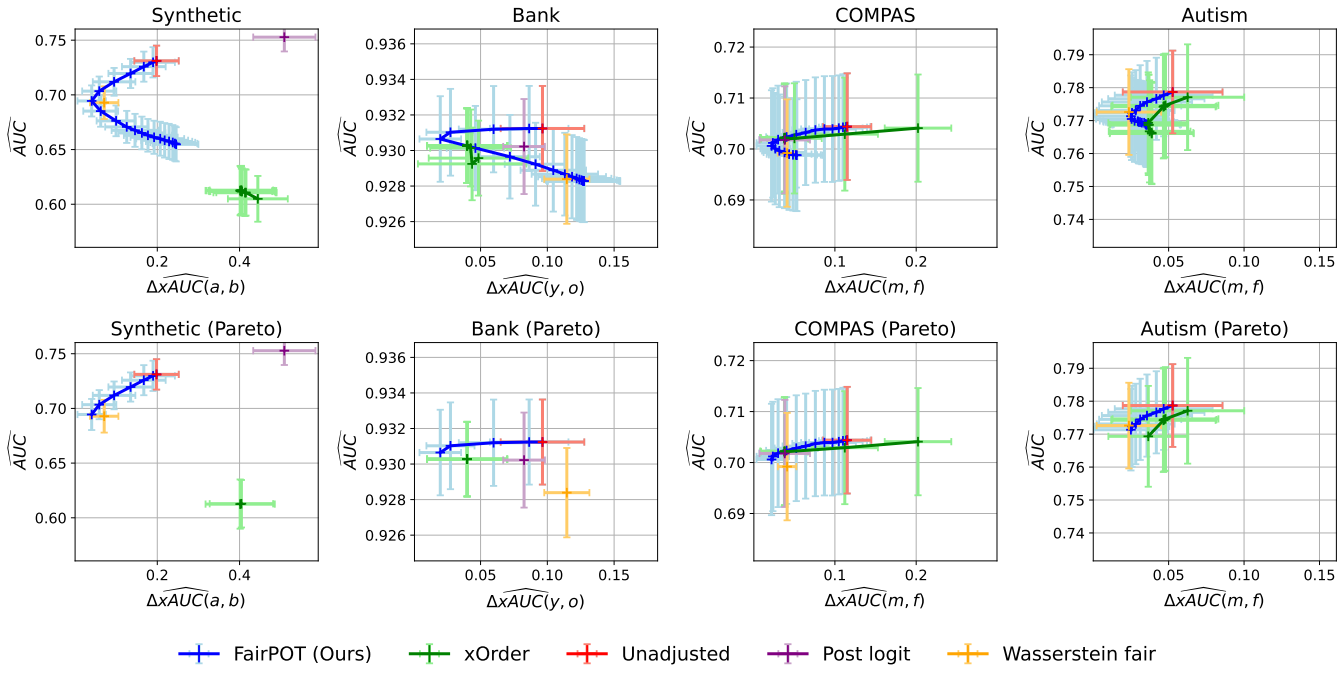

Figure 6: $\widehat{\mathrm{AUC}} - \Delta\widehat{\mathrm{xAUC}}$ trade-off (mean with std bar).

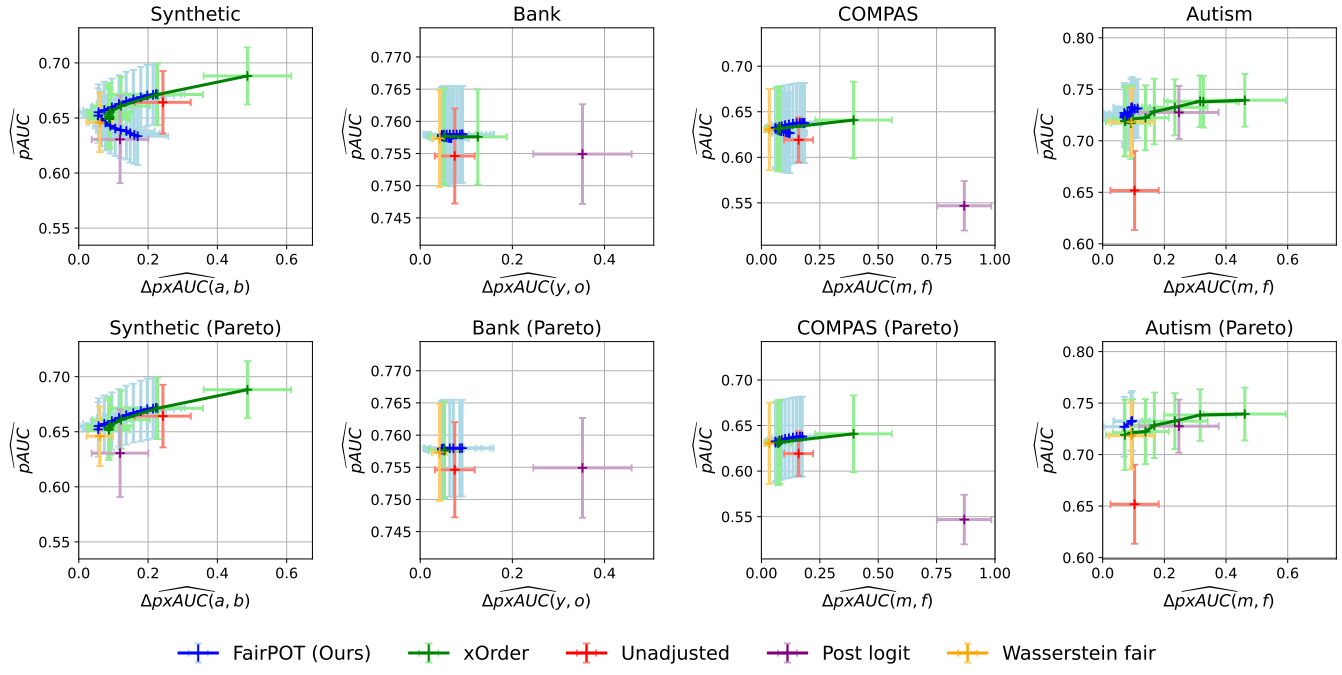

Figure 7: $\widehat{p\mathrm{AUC}} - \Delta\widehat{p\mathrm{xAUC}}$ trade-off (mean with std bar). We set $\alpha = 0.3$ for Synthetic, Bank, and COMPAS datasets, and $\alpha = 0.1$ for Autism.

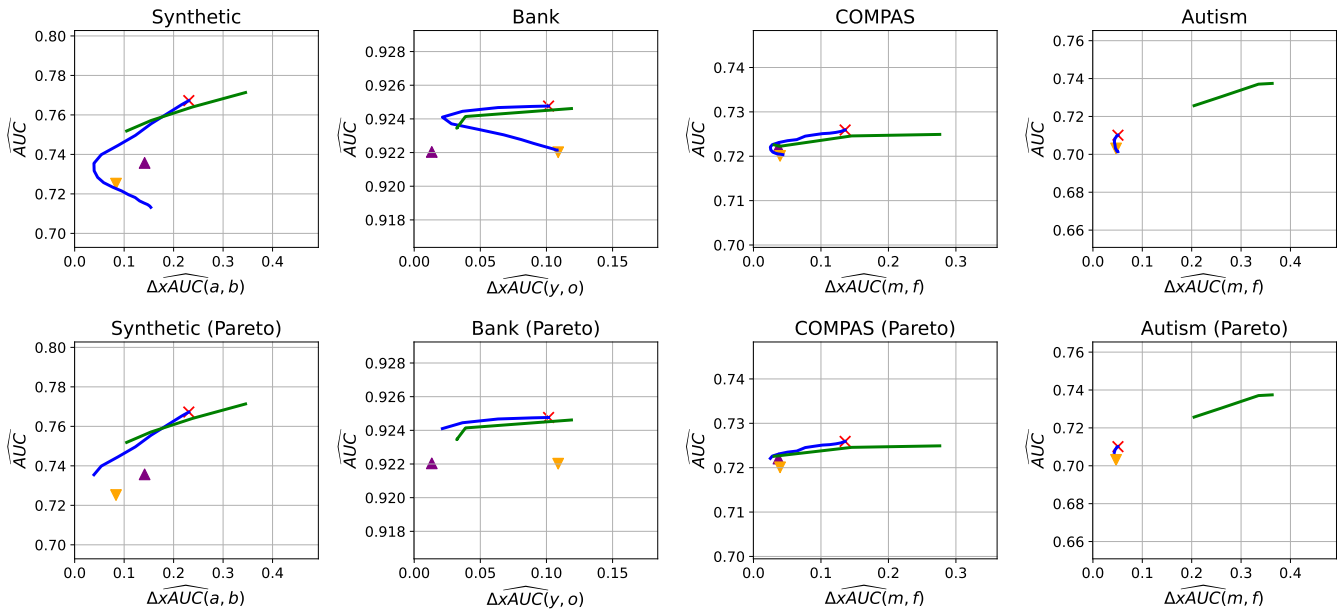

Figure 8: $\widehat{\text{AUC}} - \Delta\widehat{\text{xAUC}}$ trade-off (Gradient Boosting).

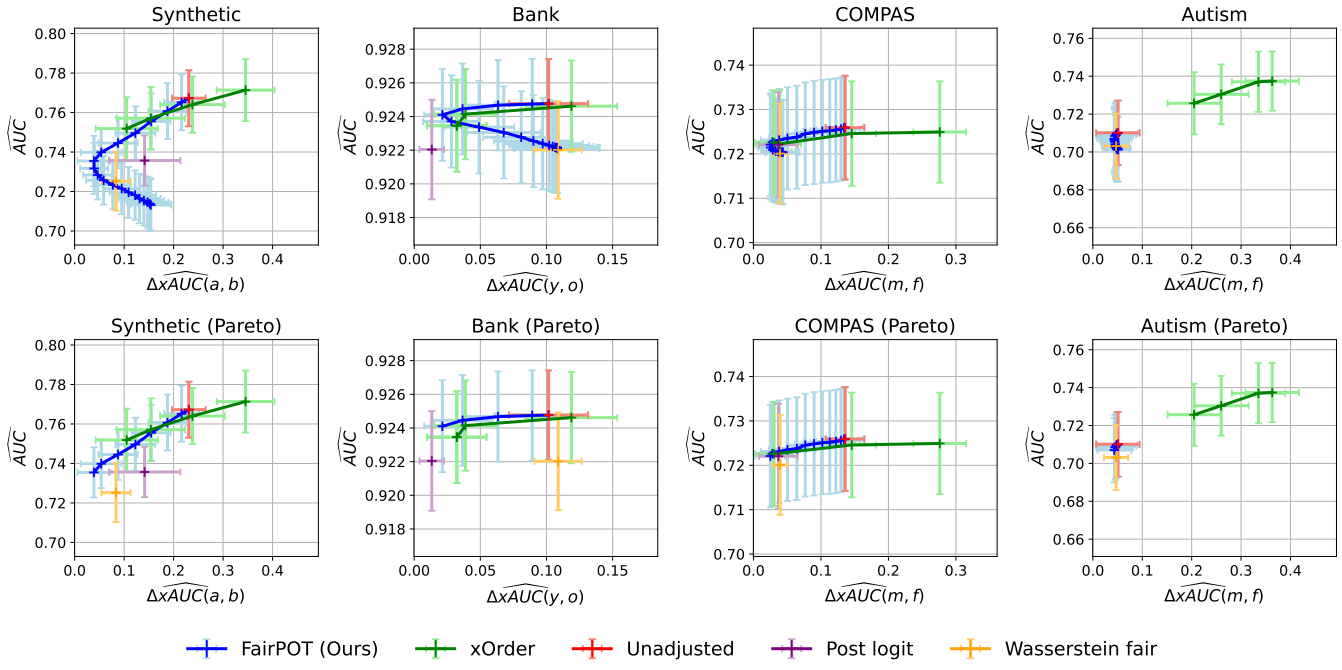

Figure 9: $\widehat{\text{AUC}} - \Delta\widehat{\text{xAUC}}$ trade-off (Gradient Boosting; mean with std bar).

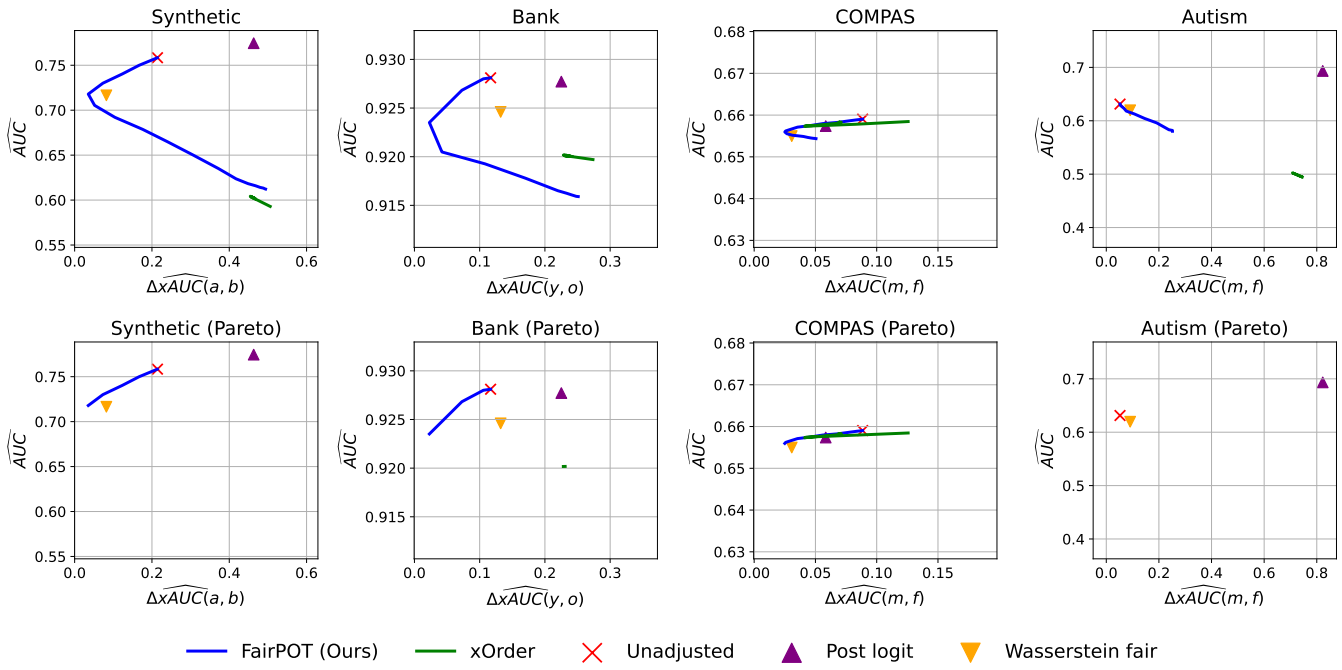

Figure 10: $\widehat{\text{AUC}} - \Delta\widehat{\text{xAUC}}$ trade-off (Random Forest).

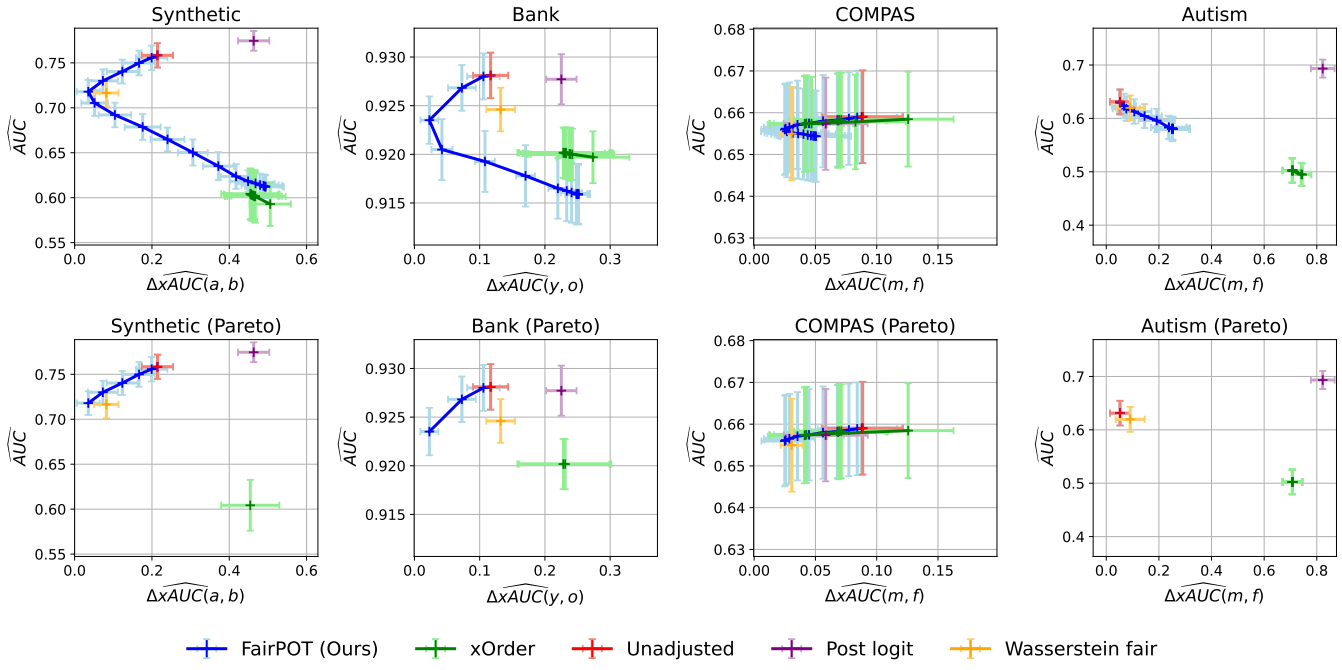

Figure 11: $\widehat{\text{AUC}} - \Delta\widehat{\text{xAUC}}$ trade-off (Random Forest; mean with std bar).

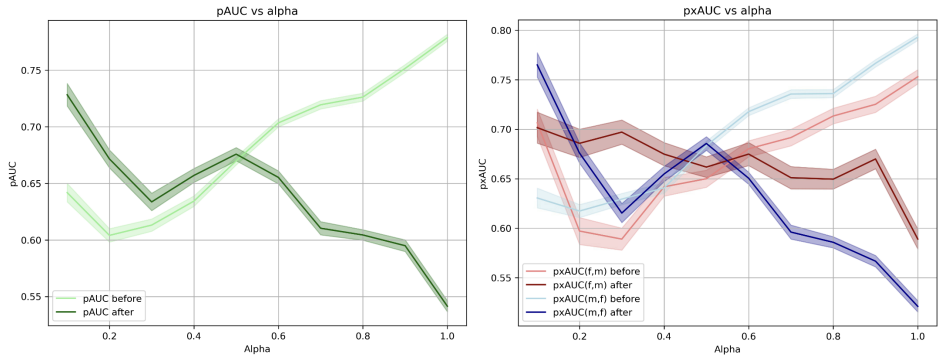

Figure 12: pAUC vs α, pxAUC vs α for Autism Data.

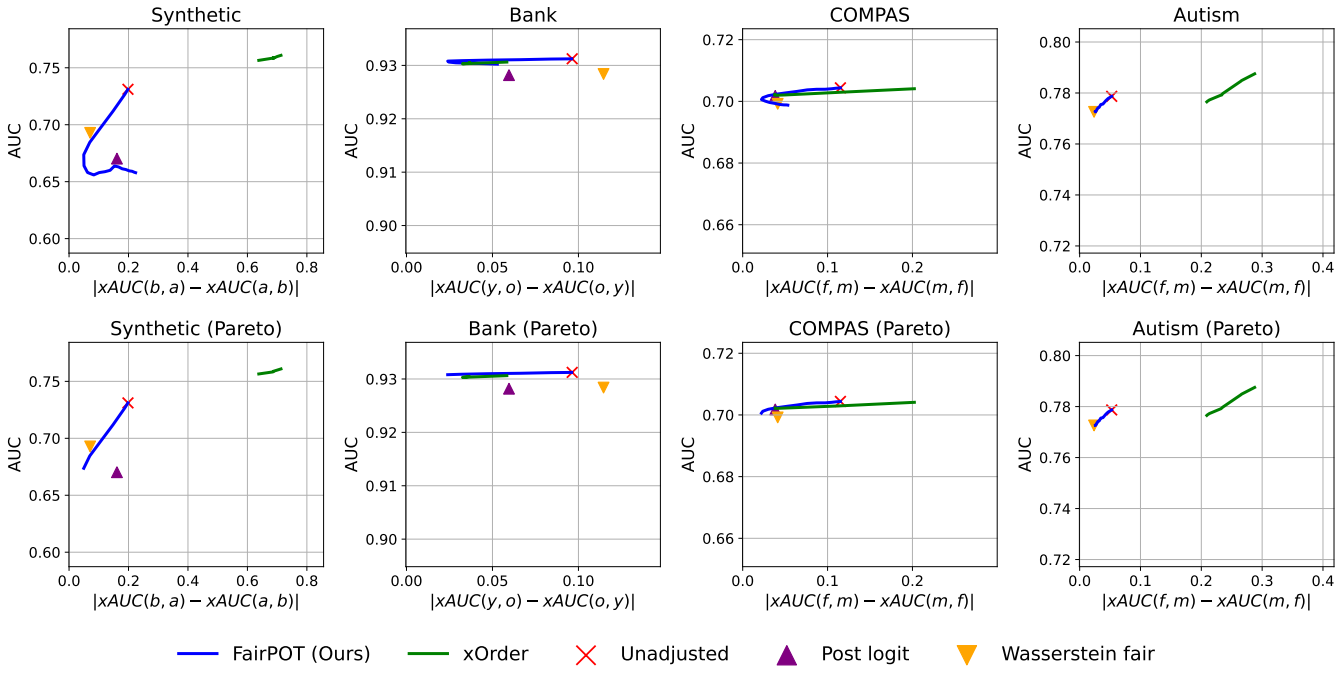

Figure 13: $\widehat{\text{AUC}} - \Delta\widehat{\text{xAUC}}$ trade-off for ablation study.

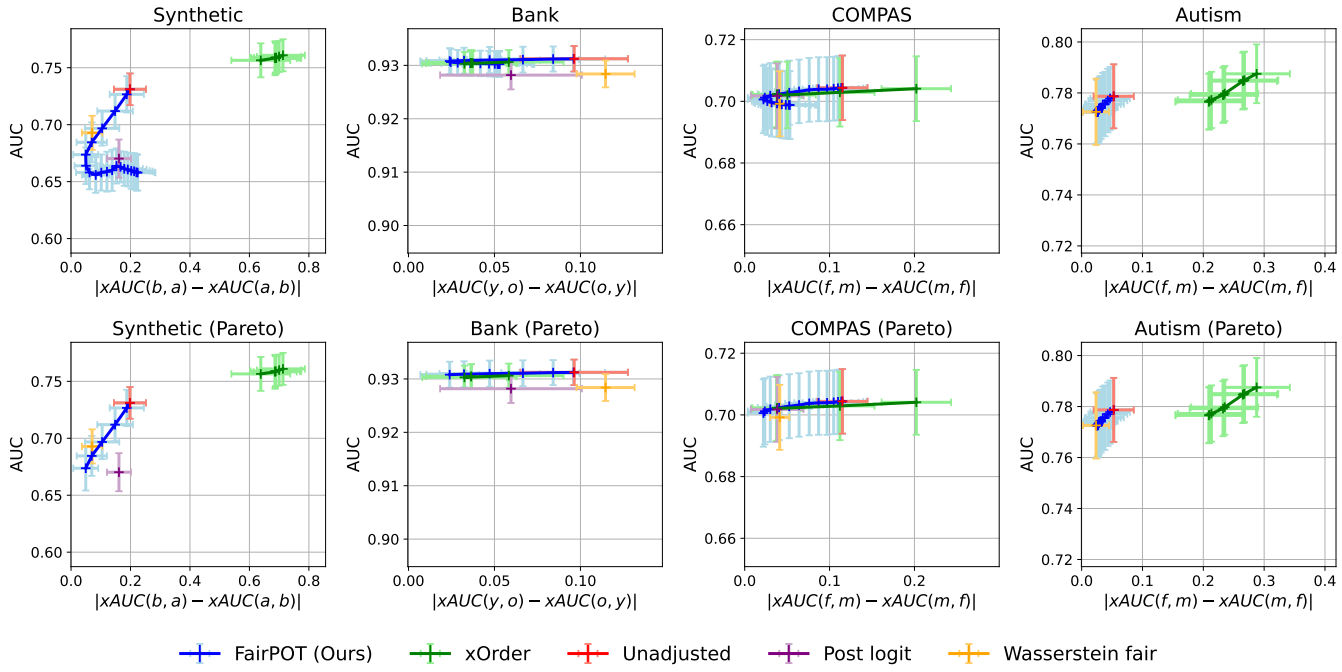

Figure 14: $\widehat{\mathrm{AUC}} - \Delta\widehat{\mathrm{xAUC}}$ trade-off for ablation study (mean with std bar).