Group-Aware Threshold Adaptation for Fair Classification

Taeuk Jang[1], **Pengyi Shi**[2], **Xiaoqian Wang**[1*]

[1]School of Electrical and Computer Engineering, Purdue University, West Lafayette, USA, 47907
[2]Krannert School of Management, Purdue University, West Lafayette, USA, 47907
{jang141, shi178, joywang}@purdue.edu

Abstract

The fairness in machine learning is getting increasing attention, as its applications in different fields continue to expand and diversify. To mitigate the discriminated model behaviors between different demographic groups, we introduce a novel post-processing method to optimize over multiple fairness constraints through group-aware threshold adaptation. We propose to learn adaptive classification thresholds for each demographic group by optimizing the confusion matrix estimated from the probability distribution of a classification model output. As we only need an estimated probability distribution of model output instead of the classification model structure, our post-processing model can be applied to a wide range of classification models and improve fairness in a model-agnostic manner and ensure privacy. This even allows us to post-process existing fairness methods to further improve the trade-off between accuracy and fairness. Moreover, our model has low computational cost. We provide rigorous theoretical analysis on the convergence of our optimization algorithm and the trade-off between accuracy and fairness. Our method theoretically enables a better upper bound in near optimality than previous method under the same condition. Experimental results demonstrate that our method outperforms state-of-the-art methods and obtains the result that is closest to the theoretical accuracy-fairness trade-off boundary.

Introduction

Machine learning is broadening its impact in various fields including credit analysis, job screening and *etc*. Consequently, the importance of fairness in machine learning is emerging. However, recent models have been found to behave differently between demographic groups in favorable predictions. For example, it has been discovered that COMPAS, the criminal risk assessment software currently used to help pretrial release decisions, has biases between different races (Dressel and Farid 2018). Specifically, blacks got higher risk scores predicted from the model than whites with similar profiles. Therefore, discrimination truly exists and resolving it is critical as its direct and potential impact is growing tremendously.

*Corresponding author.

However, obtaining fairness is not a trivial problem, as the dataset itself will be biased when it is accumulated artificially (Jang, Zheng, and Wang 2021). Simply modifying sensitive features (such as *race*, *gender*) from the data does not solve the bias, because there is indirect discrimination (Pedreshi, Ruggieri, and Turini 2008) caused by the feature relevance, which means sensitive information can be inferred from other features.

In order to alleviate discrimination from different perspectives, various quantitative measurements of group equity (Hardt, Price, and Srebro 2016; Kleinberg, Mullainathan, and Raghavan 2016; Chouldechova 2017) have been proposed. It has been proven that the pursuit of fairness is subject to a trade-off between fairness and accuracy (Liu et al. (2019), Kim et al. (2020)).

Moreover, Pleiss et al. (2017) studied the trade-offs between fairness notions that cannot be satisfied at the same time. Therefore, recent works (Feldman et al. 2015; Zhang, Lemoine, and Mitchell 2018; Hardt, Price, and Srebro 2016) usually target at a certain fairness notion. However, these approaches suffer from the *lack of flexibility*, *i.e.,* target fairness cannot be adjusted to meet the needs. If the fairness constraints change under some circumstances, traditional fairness models need to be re-trained from scratch, which is computationally demanding and sometimes inapplicable due to model settings.

To overcome the limitations, we propose a novel post-processing method to improve fairness in a model-agnostic manner *i.e.,* we only need the prediction of an unknown model. Our GSTAR (Group Specific Threshold Adaptation for faiR classification) model learns adaptive classification thresholds for each demographic group in classification task to improve the trade-off between fairness and accuracy. Given an existing classification model, GSTAR approximates the probability distribution of the model output and utilizes confusion matrix to quantify accuracy and fairness w.r.t. the group-aware classification thresholds. This allows us to: 1) prevent from burdening additional complexity or deteriorate the stability of the training process of the classifier; 2) integrate different fairness notions into one unified objective function; 3) easily adapt one pre-trained model to other fairness constraints.

We summarize our contributions of this paper as follows:

1. We propose a novel post-processing method, named

GSTAR, which can learn group-aware thresholds to optimize the fairness-accuracy trade-off in classification. We empirically show that GSTAR outperforms state-of-the-art methods.

2. With GSTAR, we can simultaneously optimize over multiple fairness constraints with a low computational cost. GSTAR does not require multiple iterations over data, instead, it takes *at most* one pass of data in training for fast computation.
3. We conduct extensive rigorous theoretical analysis on our method, in terms of convergence analysis and fairness-accuracy trade-off. We introduce theoretical improvement in terms of near optimality.
4. We derive Pareto frontiers of our model for the fairness-accuracy trade-offs that contextualize the quality of fair classification.

Related Works

In order to achieve group fairness, which quantifies the discrimination among different sensitive groups, a diverse notion of fairness has been introduced. Equalized odds (Hardt, Price, and Srebro 2016) enforce equality of true positive rates and false positive rates between different demographic groups. Pleiss et al. (2017) relaxed equalized odds to satisfy group-wise calibration. Demographic parity or disparate impact (Barocas and Selbst 2016) suggests that a model is unbiased if the model prediction is independent of the protected attribute.

Among different fairness methods, post-processing techniques propose to improve fairness by modifying the output of a given classifier. Hardt et al. (2016) propose to ensure equalized odds by constraining the model output. Kim et al. (2020) utilize confusion matrix and propose least-square accuracy-fairness optimization problem. Kamiran et al. (2012) propose to give a favorable outcome to unprivileged and an unfavorable outcome to the privileged group when the confidence of the prediction is in a certain range. However, such *static* confidence window keeps the same regardless of the demographic group and is determined by grid search, so it is less efficient.

Threshold adjustment (a.k.a. thresholding) was introduced to improve the performance of *static* thresholds. In the literature, Menon et al. (2018) prove that instance-dependent thresholding of the predictive probability function is the optimal classifier in cost-sensitive fairness measures. Also, when considering immediate utility, Corbett-Davies et al. (2017) show that optimal algorithm is achieved from group-specific threshold which is determined by group statistics. However, to the best of our knowledge, the threshold adjustment approach has not been deeply studied that neither encompasses broad group fairness metrics nor describes an explicit method to achieve the threshold.

Trade-off between fairness and accuracy exists when we impose fairness constraint to a model. Recent studies (Chouldechova 2017; Zhao and Gordon 2019) prove that models targeting at such fairness notions conform to an information theoretic lower bound on the joint error across different sensitive groups. Therefore, our work presents a practical upper bound of the best achievable accuracy given the fairness constraints.

Here, our work is the most related to the post-processing methods (Hardt, Price, and Srebro 2016; Kim, Chen, and Talwalkar 2020). However, ours differ from theirs in several aspects. First, we theoretically prove that GSTAR achieves a better upper bound of near optimality than Hardt et al. (2016) as we directly operate on ROC curve instead of linear intersections in Hardt et al. (2016). Also, GSTAR corrects the predicted label by the confidence of the prediction from a given model instead of randomly flipping the output to achieve equalized odds, which is more reliable in post-processing. FACT (Kim, Chen, and Talwalkar 2020) utilizes a single point (static) from the classifier to be post-processed as a reference which does not fully utilize the classifier for the post-processing. In contrast, by approximating the distribution of the continuous predicted logits, GSTAR model enables a larger feasible region than Kim et al. (2020) with a better fairness-accuracy trade-off. We validate the improvement in this trade-off via both theoretical and empirical results. It is notable that these related methods can be considered as a special case of GSTAR.

GSTAR for Fair Classification

Motivation

Consider a binary classification problem with a binary sensitive feature, such that the sensitive feature $A \in \{0, 1\}$ and label $Y \in \{0, 1\}$. In general, for a given data X, a binary classification model outputs an unnormalized logit $h(X) \in \mathbb{R}$ with the class label probability $R(X) = \sigma(h(X)) \in [0, 1]$, where σ is an activation function (*e.g.,* sigmoid function). It is not necessary to calculate R in a classification model, *e.g.,* support vector machines directly use the positiveness/negativeness of logit $h(X)$ to determine classification outcome.

For traditional models, we use a cut-off threshold $\theta_h = 0$ for $h(X)$ (*i.e.,* $\theta_R = \sigma(0) = 0.5$ for $R(X)$) in classification, such that the predicted label is determined by $\hat{Y} = \mathbb{I}\{h(X) \geq \theta_h\}$. In the following context, unless otherwise mentioned, we use θ to refer to the threshold θ_h on logit h since it is applicable to a wider range of classification models, and the corresponding threshold on label probability θ_R can be easily inferred from the threshold on logit h. Traditional models use the same cut-off threshold θ for different demographic groups. However, since the distribution of logits h in different demographic groups can be different, using the same threshold θ brings biased classification.

In Fig. 1, we use a real-world example of image classification on CelebA dataset with ResNet50 (He et al. 2016) to show that the default setting of classification thresholds affects both accuracy and fairness in classification. The goal is to predict whether the image of a person is attractive or not, and consider sensitive attribute as gender. This can be generalized to different sensitive attributes in image classification task, *e.g.,* age or race (Lokhande et al. 2020). We can observe an obvious difference in the distribution of logit h between two gender groups. If we use a unified classification threshold $\theta_1 = \theta_0 = 0$, it naturally brings a difference in the true positive rate and true negative rate between two

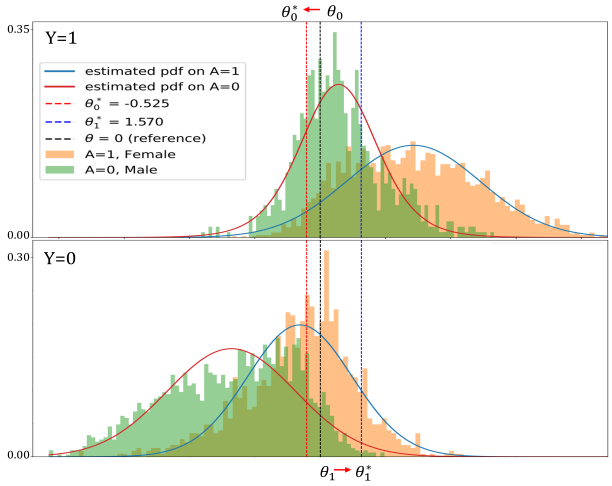


Figure 1: Histograms of logit h distribution from logistic regression on CelebA data, where θ is the threshold to assign predicted label based on h. The top and bottom plot is for positive samples ($Y = 1$, attractive), and negative samples ($Y = 0$, unattractive). Bars represent the distributions of logit h of sensitive groups, and curves are estimated probability density functions of logit h of sensitive groups as in the legend. $\theta = 0$ (black dashed line) is the default classification thresholds. The default thresholds result in biased prediction towards the unprivileged group ($A = 0$) due to the different logit h distributions in different sensitive groups. (θ_0^*, θ_1^*) (colored dashed line) are group-aware thresholds for each sensitive group achieved by GSTAR.

gender groups, thus it behaves as a biased classification. Instead, we observe that the optimal group-specific threshold obtained from GSTAR ($\theta_1^* > \theta_1$, and $\theta_0^* < \theta_0$) can adapt to such discrepancy in distribution between two demographic groups to improve both fairness and accuracy.

Group-Aware Classification Thresholds

Given an existing classification model and a sensitive attribute a, we can denote true positive rate (TP_a), false positive rate (FP_a), true negative rate (TN_a), and false negative rate (FN_a) in the confusion matrix. Most fairness notions can be represented with entries in the confusion matrix. For instance, Equal Opportunity (EOp) (Hardt, Price, and Srebro 2016) requires $TP_0 = TP_1$, and Demographic Parity (DP) (Barocas and Selbst 2016) requires

$$\frac{TP_1 n_{11} + FP_1 n_{01}}{N_1} = \frac{TP_0 n_{10} + FP_0 n_{00}}{N_0},$$

where n_{ya} denotes the number of samples in the subset $\{Y = y, A = a\}$, $N_a = \sum_y n_{ya}$ denotes the number of samples in $\{Y = y\}$, and $N = \sum_{y,a} n_{ya}$ is the total number of samples.

Consider the group-aware classification threshold $\boldsymbol{\theta} = (\theta_1, \theta_0)^\mathsf{T}$, where θ_a is the classification threshold for sensitive group $A = a$. We can formulate the entries in the confusion matrix w.r.t. $\boldsymbol{\theta}$ as below:

$$\begin{aligned} \text{TP}_a(\theta_a) &\approx 1 - \int_{-\infty}^{\theta_a} f_{1a}(x)dx, \quad \text{FN}_a(\theta_a) \approx 1 - \text{TP}_a(\theta_a) \\ \text{FP}_a(\theta_a) &\approx 1 - \int_{-\infty}^{\theta_a} f_{0a}(x)dx, \quad \text{TN}_a(\theta_a) \approx 1 - \text{FP}_a(\theta_a) \end{aligned} \tag{1}$$

where $f_{ya}(x)$ is an estimated probability density function of the distribution of output logit h in the subset $\{Y = y, A = a\}$.

Then, we formulate the fairness-constrained classification problem with the objective of minimizing classification error into a least-squared optimization problem. We denote our objective function as $\mathcal{L}(\boldsymbol{\theta})$ which consists of the performance loss $\mathcal{L}_{per}(\boldsymbol{\theta})$ and fairness loss $\mathcal{L}_{fair}(\boldsymbol{\theta})$ that are represented with the entries of the confusion matrix. In other words, our goal is to minimize the objective function $\mathcal{L}(\boldsymbol{\theta})$ as below:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_{per}(\boldsymbol{\theta}) + \lambda \mathcal{L}_{fair}(\boldsymbol{\theta}), \tag{2}$$

where λ is a hyperparameter that determines how much fairness is enforced in the optimization. The performance error $\mathcal{L}_{per}(\boldsymbol{\theta})$ can be written as

$$\begin{aligned} \mathcal{L}_{per}(\boldsymbol{\theta}) = \Big(&\frac{n_{01}}{N}\text{FP}_1(\theta_1) + \frac{n_{11}}{N}\text{FN}_1(\theta_1) \\ &+ \frac{n_{00}}{N}\text{FP}_0(\theta_0) + \frac{n_{10}}{N}\text{FN}_0(\theta_0) \Big)^2. \end{aligned}$$

As for $\mathcal{L}_{fair}(\boldsymbol{\theta})$, it can be formulated to any fairness metrics that are expressible with confusion matrix. For instance, when we impose EOp ($\text{TP}_1 = \text{TP}_0$) and predictive equality (PE) ($\text{FP}_1 = \text{FP}_0$) (Chouldechova 2017), we can get the corresponding $\mathcal{L}_{fair}(\boldsymbol{\theta})$ by summing over the least squared form of each constraint. Also, satisfying EOp and PP is equivalent to satisfying Equalized Odds (EOd) (Hardt, Price, and Srebro 2016), This can be formulated in our $\mathcal{L}_{fair}$ as

$$\begin{aligned} \mathcal{L}_{fair}^{EOd}(\boldsymbol{\theta}) &= \mathcal{L}_{fair}^{EOp}(\boldsymbol{\theta}) + \mathcal{L}_{fair}^{PP}(\boldsymbol{\theta}) \\ &= \big(TP_1(\theta_1) - TP_0(\theta_0)\big)^2 + \big(FP_1(\theta_1) - FP_0(\theta_0)\big)^2. \end{aligned} \tag{3}$$

Note that a lower $\mathcal{L}_{fair}$ value indicates a fairer threshold. When $\mathcal{L}_{fair}^{EOD}(\boldsymbol{\theta}) = 0$, we can interpret as the $\boldsymbol{\theta}$ satisfies the perfect EOd fairness. Similar to (3), we can enforce multiple fairness constraints by summing over the least square of each metric with different weight constant λ to each fairness constraints if needed.

Also, it is notable that compared to FACT (Kim, Chen, and Talwalkar 2020) that enforces fairness through confusion tensor, our formulation of fairness in $\mathcal{L}_{fair}(\boldsymbol{\theta})$ represents a direct notion of fairness metrics and improves the measures that allows us to achieve better performance and Pareto frontiers that is shown in Section and Fig. 2. For example, FACT integrates multiple constraints as a weighted sum with the weights being the number of samples in each class. In this expression, the imbalance between the two fairness criteria will grow as the degree of imbalance in the data increases. In contrast, our formulation expresses the constraints as the exact notion of each metric that is not biased by the statistics of the datset and we observe improved Pareto frontier as in Fig. 2.

Optimization of GSTAR

Our GSTAR objective in (2) lies in the family of Nonlinear Least Squares Problem (NLSP) (Gratton, Lawless, and Nichols 2007). To optimize objective (2) and find the threshold $\boldsymbol{\theta}$, we adopt the Gaussian-Netwon optimization method (Gratton, Lawless, and Nichols 2007). Here we take EOp constraint as an example to show the alternating optimization steps, then $\mathcal{L}_{fair}(\boldsymbol{\theta})$ can be written as

$$\mathcal{L}_{fair}^{EOp}(\boldsymbol{\theta}) = \left(\text{TP}_1(\theta_1) - \text{TP}_0(\theta_0)\right)^2. \tag{4}$$

To solve NLSP with the Gauss-Newton method, we first convert the nonlinear optimization problem to a linear least square problem using Taylor expansion. That is, the parameter values are calculated in an iterative fashion with

$$\theta_a \approx \theta_a^{k+1} = \theta_a^k + \Delta_a, \tag{5}$$

in the k-th iteration number, with the vector of increments $\Delta = \{\Delta_a\} = \{\theta_a^{k+1} - \theta_a^k\}$ (also known as the shift vector).

We rewrite our objective function as a real vector function $r(\boldsymbol{\theta}) = \big(r_1(\boldsymbol{\theta}), r_2(\boldsymbol{\theta})\big) = (\mathcal{L}_{per}, \lambda\mathcal{L}_{fair})$. We linearize each component in the loss function to a first-order Taylor polynomial expansion as

$$r_i(\boldsymbol{\theta}) \approx r_i(\boldsymbol{\theta}^k) + \sum_a \frac{\partial r_i(\boldsymbol{\theta}^k)}{\partial \theta_a}\Delta_a \tag{6}$$

with $\boldsymbol{\theta}^k = (\theta_0^k, \theta_1^k)$. Plugging this linearized equation into the objective function, we get the usual least square problem. Then, the optimal solution can be obtained as

$$\Delta = -(J^T J)^{-1} J^T f(\boldsymbol{\theta}^k), \tag{7}$$

where $J = \{J_{ia}\}$ with $J_{ia} = \{\frac{\partial r_i(\boldsymbol{\theta})}{\partial \theta_a}\}$ is the Jacobian. Each entry of the jacobian can be expressed with linear combination of pdf and cdf of f_{ya} for $i, a, y \in \{0, 1\}$. we can finalize the alternating optimization as

$$\theta_0^\tau = \theta_0^{\tau-1} + \Delta_0^\tau, \quad \theta_1^\tau = \theta_1^{\tau-1} + \Delta_1^\tau. \tag{8}$$

It is notable that in each iteration we derive the optimal update step Δ_a, which eliminates the burden of tuning hyperparameter (such as learning rate) in iterative algorithm. See the supplementary for detailed optimization process.

The alternating optimization of GSTAR model is of low computational cost. We take at most one pass of the data for learning the estimated probability density functions f_{ya} in (1) (we do not even need to traverse the data if the parameters (such mean and variance in Gaussian distribution) for the estimated probability density functions f_{ya} can be provided). The optimization of $\boldsymbol{\theta}$ with alternating optimization is efficient since we only need f_{ya}. Therefore, we need a constant time for each update. Overall, the time complexity of GSTAR is $O(n + T)$, where n is the number of samples, and T is the number of iterations in alternating optimization.

Besides, if a unified threshold is necessary (Corbett-Davies et al. 2017), *i.e.,* $\theta_1 = \theta_0$, the optimization algorithm also applies and we only have one scalar variable in (2). When we have a unified threshold, we do not require sensitive information in the testing phase that we can conform more strict privacy regulations than group-aware thresholding. However, we have to sacrifice both fairness and accuracy as the thresholding is less flexible.

Theoretical Analysis

Upper Bounds on FPR/FNR Gap between Groups We first state the assumptions we need to make for Theorem 1 and 2.

Assumption 1 *For any given classier h and its induced PDF f_{ya} and CDF F_{ya}, we assume the following holds:*

- *The PDF $f_{ya}(x)$ is uniformly bounded, i.e., there is an $\hat{f}_{ya}(x) = \max_x f_{ya}(x)$.*
- *The inverse CDF $F_{ya}^{-1}(x)$ is Lipschitz continuous with Lipschitz constant M_{ya}.*
- *The difference in the CDF between two groups is uniformly bounded, i.e.,*

$$|F_{y1}(x) - F_{y0}(x)| \le u_y, \ \ \forall x.$$

Theorem 1 *For any given classifier that satisfies Assumption 1 and any given pair of thresholds (θ_0, θ_1) that satisfies the perfect EOp condition, the gap between false-positive rates of the two group is upper bounded by*

$$|\epsilon_1| = \big|FP_0(\theta_0) - FP_1(\theta_1)\big| \le u_0 + C_1 u_1, \tag{9}$$

where $C_1 = \hat{f}_{01} M_{10}$.

Theorem 2 *For any given classifier that satisfies Assumption 1 and any given pair of thresholds (θ_0, θ_1) that satisfies the perfect PE condition, the gap between false-negative rates of the two group is upper bounded by*

$$|\epsilon_2| = \big|FN_0(\theta_0) - FN_1(\theta_1)\big| \le u_1 + C_0 u_0, \tag{10}$$

where $C_0 = \hat{f}_{11} M_{00}$.

Theorem 1 and 2 characterize the upper bound of false positive/negative rate gap between two groups when the false negative/positive rate gap is 0. At the same time, it captures the upper bound of additional accuracy loss due to the two different thresholds for different groups under a perfect fairness (EOp or PE) condition.

Trade-off between Accuracy and Fairness Now we prove a theorem to characterize the trade-off between accuracy and fairness. Let $\theta_a^* = \text{argmin}_{\theta_a} \mathcal{L}_{per}(\theta_a)$, and its perturbed value $\tilde{\theta}_a$ as

$$\begin{aligned} |\text{FN}_1(\theta_1{}^*) - \text{FN}_1(\tilde{\theta}_1)| &\le \gamma/2, \\ |\text{FN}_0(\theta_0{}^*) - \text{FN}_0(\tilde{\theta}_0)| &\le \gamma/2, \end{aligned} \tag{11}$$

for some perturbation coefficient γ. Then for optimal perturbed version $\tilde{\theta}_a^* = \text{argmin}_{\tilde{\theta}_a} \mathcal{L}_{per}(\tilde{\theta}_a)$, we state the theorem below:

Theorem 3 *Under Assumption 1 and condition* (11)*,*

$$\mathcal{L}_{per}(\theta_1^*) - \mathcal{L}_{per}(\tilde{\theta}_1^*) \quad \le \quad C\gamma,$$

where

$$C = 2L^*\left(\frac{r_1}{2} + r_0\frac{\hat{f}_{01}M_{11}}{2} + \frac{n_{00}}{N}\left(\hat{f}_{00}M_{10} + \frac{\hat{\epsilon_1'}M_{11}}{2}\right) + \frac{n_{10}}{N}\right)$$

and $\hat{\epsilon}_1' = \max \tilde{\epsilon}_1'$ is the maximum of the derivative of $\tilde{\epsilon}_1$.

Theorem 3 quantifies the decrease in accuracy loss (*i.e.,* the improvement in accuracy) when we allow a gap of true positive rates between two groups, i.e., relaxation from the perfect fairness cases in Theorem 1 and 2.

Convergence Analysis of GSTAR Our objective function and the optimization solution algorithm belong to the family of Gauss-Newton algorithm. Given the assumptions A1 and A2 below,

- A1. There exists θ^* such that $J^T(\theta^*)r(\theta^*) = 0$,
- A2. The Jacobian at θ^* has full rank,

we state the following theorem of convergence:

Theorem 4 *Assume that the estimated density function $f(\cdot)$ satisfy assumptions A1 and A2. Further, $f(\cdot)$ satisfies that*

$$||Q(\theta^k)(J^T J)^{-1}(\theta^k)||_2 \leq \eta$$

for some constant $\eta \in [0, 1)$ for each iteration k, where $Q(\theta)$ denotes the second order terms $\sum_i r_i(\theta)\nabla^2 r_i(\theta)$. Then as long as the initial solution is sufficiently close to the true optimal with $||\theta^0 - \theta^*||_2 \leq \epsilon$, the sequence of Gauss-Newton iterates $\{\theta^k\}$ converges to θ^*.

Near Optimality of GSTAR Following the proof of Theorem 5.6 of Hardt et al. (2016), we provide the following near optimality theorem for our GSTAR model.

Theorem 5 *With a bounded loss function ℓ and a given estimated density function $f(x)$, let $\hat{R}_h \in [0, 1]$ be the induced random variable from the density $f(x)$ of logit $h(x)$. Then the equalized odds predictor $\hat{Y}_h$ derived from $(\hat{R}_h, A)$ using the method in our paper can achieve near optimality in the following sense:*

$$\mathbb{E}[\ell(\hat{Y}_h, Y)] \leq \mathbb{E}[\ell(Y^*, Y)] + 2d_K(\hat{R}_h, R^*).$$

Here, Y is the true label, Y^* is the optimal equalized odds predictor derived from the Bayes optimal regressor R^* as given in Hardt et al. (Hardt, Price, and Srebro 2016), and $d_K(\hat{R}_h, R^*)$ is the conditional Kolmogorov distance.

Theorem 5 provides that GSTAR has tighter bound of near optimality than Hardt et al. (2016) under the same condition. See the supplementary for the proof of Theorem 1 - 5.

Experiments

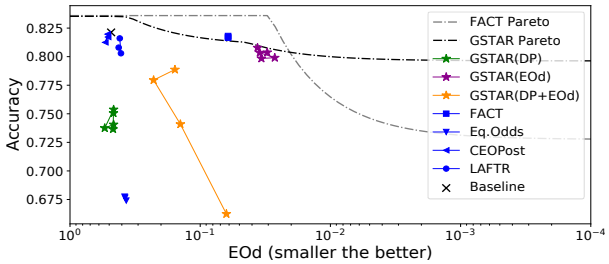
In this section, we validate GSTAR model on four well-known fairness datasets and compare with other state-of-the-art methods.

Experimental Setup

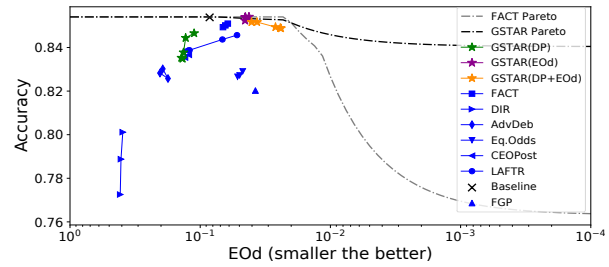
We compare with multiple fairness approaches in the experiments. For clear demonstration of results, we use different shapes of marker for each comparing methods in Fig. 2 and Fig. 4. The comparing methods include: FGP (Tan et al. 2020), FACT (Kim, Chen, and Talwalkar 2020), DIR (Feldman et al. 2015), AdvDeb (Zhang, Lemoine, and Mitchell 2018), CEOPost (Pleiss et al. 2017), Eq.Odds (Hardt, Price, and Srebro 2016), LAFTR (Madras et al. 2018), and Baseline: For CelebA dataset, we use ResNet50 (He et al. 2016) as a reference, and logistic regression for all other datasets.

We choose broadly used fairness metrics in evaluation including: equal opportunity difference (EOp) and equalized odds difference (EOd) (Hardt, Price, and Srebro 2016); 1-disparate impact (1-DIMP) (Barocas and Selbst 2016); balanced accuracy difference (BD). We use balanced accuracy (BA) and accuracy (ACC) as performance metrics.

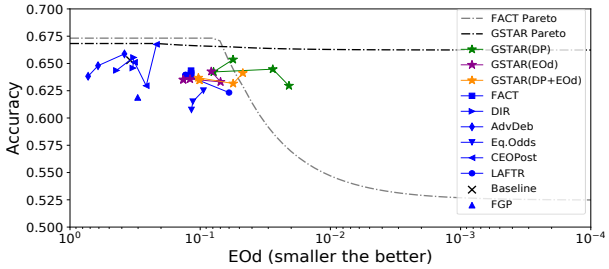
We evaluate the methods on four fairness datasets:



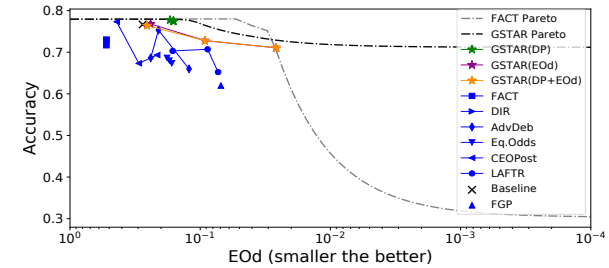
(a) CelebA Dataset



(b) Adult Dataset

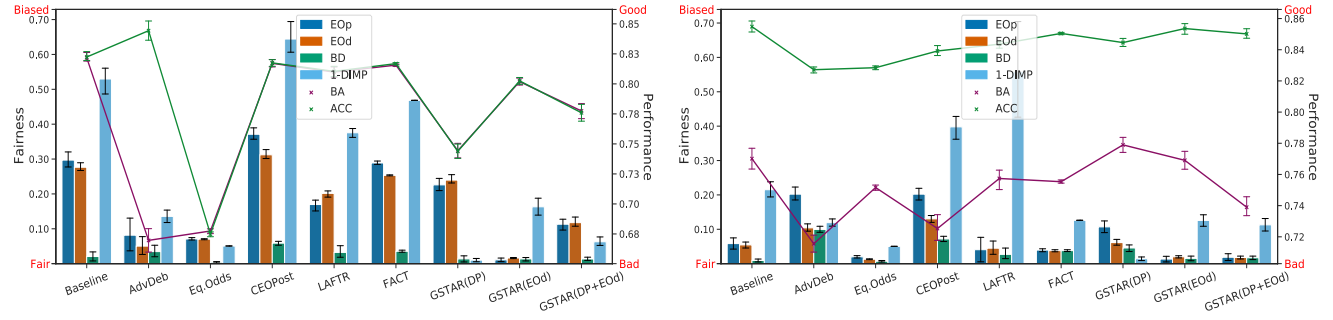


(c) Compas Dataset



(d) German Dataset

Figure 2: Pareto frontiers of equalized odds to show the upper bound of best achievable accuracy under different fairness constraints. Upper right region under the boundary is desired. The variations of GSTAR generally achieve the best trade-offs as they are the closest to the Pareto frontier.



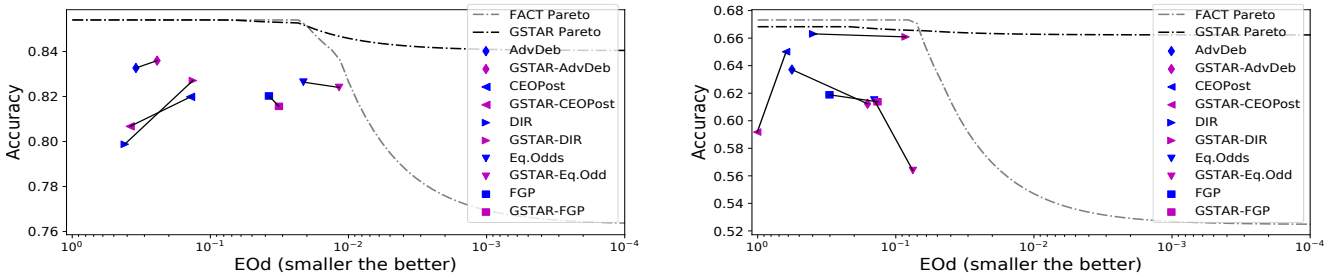
(a) CelebA Dataset

(b) Adult Dataset

(c) Compas Dataset

(d) German Dataset

Figure 3: Evaluation on fairness and performance metrics. The bar plots indicate fairness measures of each model. The line plots indicate the performance measure of each model. Lower fairness values (left y-axis) and higher performance values (right y-axis) show better fairness and performance respectively. We consider three variations of GSTAR models (DP, EOd, DP+EOd).



(a) Adult Dataset

(b) Compas Dataset

Figure 4: Illustration of post-processing (magenta colored points) on existing fairness models (blue colored points). Given the outputs of each model, GSTAR efficiently improves most existing fairness models with optimized group-aware thresholds.

CelebA dataset (Liu et al. 2015), Adult dataset (Kohavi 1996), COMPAS[1] dataset, and German dataset (Dua and Graff 2019). More details of the comparing methods, evaluation metrics, and datasets are provided in the Supplementary.

Performance and Fairness-Accuracy Trade-Offs

In this subsection, we look into the performance evaluation of GSTAR comparing with other state-of-the-art methods. We consider Pareto frontier to visualize the trade-offs between fairness and accuracy to demonstrate the measure of performance.

In Fig. 2, we plot Pareto frontier, which is the upper bound for the accuracy-fairness trade-offs, desired output locates at the upper right region under the boundary which corresponds to higher values in accuracy and lower values in fairness discrepancy. With the same fairness constraints are given, we achieve a better frontier than the FACT (Kim, Chen, and Talwalkar 2020) as we equally weigh on demographic statistics and have a better feasible region. To obtain our results (star points), we first estimate the logit distribution from the output of the baseline model, and then we get optimal adaptive thresholds with corresponding fairness

[1]https://github.com/propublica/compas-analysis

metric by updating w.r.t. the objective function in (2). Here we have three combinations of fairness imposed to GSTAR: demographic parity (DP), equalized odds (EOd), and with both constraints (DP+EOd). By post-processing on a simple baseline, we achieved significantly better fairness with small or no sacrifice in accuracy. In all datasets, GSATR got competitive or better results than other state-of-the-art methods on both fairness and accuracy.

For example, we got $\boldsymbol{\theta}^*_{EOd} = (1.570, -0.525)^\mathsf{T}$ for the CelebA dataset. This shows that we have a higher threshold for the privileged group and a lower threshold for the unprivileged group. This optimal thresholding from GSTAR allows more samples from the privileged group to be correctly predicted as unattractive that would compensate for the discrimination of the original model. In other words, this improves false positive rate difference (also known as predictive equality (Chouldechova 2017)) with a huge amount from 0.235 to 0.014. Also, true positive rate difference (also known as equality of opportunity (Hardt, Price, and Srebro 2016)) got reduced from 0.282 to 0.018. It is notable that GSTAR only sacrificed 2.2% of accuracy to bring the big improvement in fairness.

Since the objective function of our model is independent to data dimensionality, our model is much more efficient especially for high dimensional data. We mostly outperform the computational cost comparing to the other methods. The comparison of computational time and auxiliary experiments can be found in the Supplementary material.

Flexibility and Multiple Fairness Constraints

Since each fairness metric has different interests, it has been theoretically proven that they cannot be perfectly satisfied all together (Pleiss et al. 2017; Chouldechova 2017; Kleinberg, Mullainathan, and Raghavan 2016). Because of this inherent trade-offs between fairness metrics, most of the recent works focus on a single metric at a time to achieve fairness. However with GSTAR, we have the flexibility to optimize on multiple fairness constraints that can be represented in the confusion matrix format. Moreover, given the estimated distribution f_{ya} of a arbitrary classification model, we can adjust the optimal $\boldsymbol{\theta}$ based on the needs by accommodating different fairness criteria.

Fig. 3 demonstrates the result of the methods with fairness metrics and accuracy trade-off evaluations. Overall, the variations of GSTAR achieve the best fairness on each target fairness while preserving the performance. For example in Fig. 3(a), GSTAR with EOd constraint has good performance in most fairness metrics with comparable accuracy (80.3%). Comparing with GSTAR (EOd), when we introduce EOd and DP together (DP+EOd), we achieve significantly better w.r.t. DP fairness with sacrificing a small amount of accuracy and EOd.

In general, by sacrificing individual fairness performance, we could introduce multiple constraints. Also, we observe that the more fairness constraints are introduced, the more accuracy is sacrificed. We empirically found that in some cases (*e.g.,* Fig. 3(c)), introducing multiple fairness is complementary to each other that improves both conditions.

Post-Processing on an Existing Fair Model

For a binary classifier that has a single fixed classification threshold (0 for out logit, and 0.5 for label probability), we can provide better trade-off between fairness and accuracy with GSTAR. Given the logit/probability in the model-agnostic manner, we can improve the fairness as illustrated in Fig. 4. In most cases, we observe improvement in fairness after GSTAR post-processing. It is also interesting to note that by optimizing the different thresholds for each protected group, we even obtain better performance on both fairness and accuracy, which indicates that the threshold optimization can not only improve fairness but also accuracy.

However, when the distribution of the logits/probability is highly extreme (such as the results of using GSTAR to post-process CEOPost), it is difficult to estimate the distribution and thus causes erroneous optimization in GSTAR. We empirically found that when the dataset is extremely imbalanced such that we do not have enough samples to estimate the logit/probability distribution, or the given classification model is too certain to the prediction that samples are concentrated to certain output, this problem arises.

Conclusion and Discussion

In this paper, we propose a group-aware threshold adaptation method (GSTAR) to post-process in model-agnostic manner and optimize over multiple fairness constraints.We directly optimize the classification threshold for each demographic group w.r.t. the classification error and multiple fairness constraints in a unified objective function, such that we can practically achieve an optimal trade-off between accuracy and fairness in fair classification. Our method is applicable to diverse notions of group fairness as the majority of fairness notions can be expressed as a linear or quadratic equation through confusion matrix. We empirically show that GSTAR is *flexible* with fairness regularization, *efficient* with low computational cost. We also notice that the adaptive thresholds benefit accuracy in some cases. GSTAR agrees to protect *privacy* such as article 17 of EU's GDPR (Regulation 2016). We only require the estimated distribution of the output from a given model i.e., our post-processing method is oblivious to features. Thus training data is no longer needed and allowed to be discarded after training the model that to be post-processed. Thus, GSTAR can be applied to relaxed scenarios where practitioners cannot access individual-level sensitive information but have estimated distributions of logits for each sensitive group.

Further, we empirically find that GSTAR is not applicable to post-process some classification models in the following situations: 1) the model does not provide logit/probability as the outcome; 2) The model provides an extreme distribution of the output logit/probability. For example, when the model is too certain about its prediction, it will be difficult to perform probability density estimation. In our future work, we will study possible strategies to solve the above limitations, and extend GSTAR to multi-class, multi-sensitive group problems and improve the fairness-accuracy trade-off in a more general scheme.

Acknowledgements

This work was partially supported by NSF IIS #1955890, Purdue's Elmore ECE Emerging Frontiers Center.

References

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

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2): 153–163.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *KDD*, 797–806.

Dressel, J.; and Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv*, 4(eaao5580): 1–5.

Dua, D.; and Graff, C. 2019. UCI Machine Learning Repository. *University of California, Irvine, School of Information and Computer Sciences*.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *KDD*, 259–268.

Gratton, S.; Lawless, A. S.; and Nichols, N. K. 2007. Approximate Gauss–Newton methods for nonlinear least squares problems. *SIAM*, 18(1): 106–132.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *NIPS*, 3315–3323.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Jang, T.; Zheng, F.; and Wang, X. 2021. Constructing a fair classifier with generated fair data. In *AAAI*, volume 35, 7908–7916.

Kamiran, F.; Karim, A.; and Zhang, X. 2012. Decision theory for discrimination-aware classification. In *ICDM*, 924–929. IEEE.

Kim, J. S.; Chen, J.; and Talwalkar, A. 2020. Model-Agnostic Characterization of Fairness Trade-offs. *arXiv preprint arXiv:2004.03424*.

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

Kohavi, R. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, volume 96, 202–207.

Liu, L. T.; Simchowitz, M.; and Hardt, M. 2019. The implicit fairness criterion of unconstrained learning. In *ICML*, 4051–4060.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *ICCV*, 3730–3738.

Lokhande, V. S.; Akash, A. K.; Ravi, S. N.; and Singh, V. 2020. FairALM: Augmented Lagrangian Method for Training Fair Models with Little Regret. In *ECCV*, 365–381. Springer.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*.

Menon, A. K.; and Williamson, R. C. 2018. The cost of fairness in binary classification. In *ACM FAccT*, 107–118.

Pedreshi, D.; Ruggieri, S.; and Turini, F. 2008. Discrimination-aware data mining. In *KDD*, 560–568.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On fairness and calibration. In *NIPS*, 5680–5689.

Regulation, G. D. P. 2016. Regulation EU 2016/679 of the European Parliament and of the Council of 27 April 2016. *OJEU*, 43–44.

Tan, Z.; Yeom, S.; Fredrikson, M.; and Talwalkar, A. 2020. Learning fair representations for kernel models. In *AISTATS*, 155–166.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *AIES*, 335–340.

Zhao, H.; and Gordon, G. 2019. Inherent tradeoffs in learning fair representations. In *NeurIPS*, 15675–15685.

Supplementary Material for "Group-Aware Threshold Adaptation for Fair Classification"

Optimization Procedure of GSTAR

The threshold $\boldsymbol{\theta}$ is optimized with alternating optimization method in GSTAR. Here we take EOp constraint as an example to show the alternating optimization steps, then $\mathcal{L}_{fair}(\boldsymbol{\theta})$ can be written as

$$\mathcal{L}_{fair}^{EOp}(\boldsymbol{\theta}) = \left(\text{TP}_1(\theta_1) - \text{TP}_0(\theta_0)\right)^2, \tag{1}$$

and overall objective is to minimize

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_{per}(\boldsymbol{\theta}) + \lambda \mathcal{L}_{fair}^{EOp}(\boldsymbol{\theta}). \tag{2}$$

The first step is to fix θ_0 and update θ_1. We can approximate the terms that are related to θ_1 (*e.g.,* $\text{TP}_1, \text{FP}_1, \text{TN}_1, \text{FN}_1$) in (9) with first-order Taylor expansion at $\theta_1^{\tau-1}$. For example,

$$\text{TP}_1(\theta_1) \approx \text{TP}_1(\theta_1^{\tau-1}) + \frac{\partial \text{TP}_1}{\partial \theta_1}\Big|_{\theta_1=\theta_1^{\tau-1}} (\theta_1 - \theta_1^{\tau-1}) \tag{3}$$

From (9), we can easily derive that

$$\begin{aligned} \text{TP}_1(\theta_1^{\tau-1}) = \quad & 1 - \int_{-\infty}^{\theta_1^{\tau-1}} f_{11}(x)dx, \\ \frac{\partial \text{TP}_1}{\partial \theta_1} = \quad & - f_{11}(\theta_1^{\tau-1}). \end{aligned} \tag{4}$$

Similarly, we can find the first order Taylor expansion of FP_1, FN_1, and TN_1. Then, the update of θ_1 w.r.t. (2) can be approximated with the following minimization problem w.r.t. Δ_1

$$\Delta_1^\tau := \underset{\Delta_1}{\text{argmin}} (\eta_1^\tau + \alpha_1^\tau \Delta_1)^2 + \lambda(\epsilon_1^\tau + \beta_1^\tau \Delta_1)^2, \tag{5}$$

where $\Delta_1 = \theta_1 - \theta_1^{\tau-1}$ and

$$\begin{aligned} \alpha_1^\tau =& \frac{n_{11}}{N} f_{11}(\theta_1^{\tau-1}) - \frac{n_{01}}{N} f_{01}(\theta_1^{\tau-1}), \beta_1^\tau = -f_{11}(\theta_1^{\tau-1}), \\ \eta_1^\tau =& \int_{-\infty}^{\theta_1^{\tau-1}} \big(\frac{n_{11}}{N} f_{11}(x) + \frac{n_{01}}{N}(1 - f_{01}(x)\big) dx \\ & + \int_{-\infty}^{\theta_0^{\tau-1}} \big(\frac{n_{10}}{N} f_{10}(x) + \frac{n_{00}}{N}(1 - f_{00}(x)\big) dx, \\ \epsilon_1^\tau =& \int_{\infty}^{\theta_1^{\tau-1}} f_{11}(x)dx - \int_{\infty}^{\theta_0^{\tau-1}} f_{10}(x)dx. \end{aligned} \tag{6}$$

Algorithm 1: Optimization Algorithm of GSTAR Model

Input dataset $\mathcal{X} \times \mathcal{A} \times \mathcal{Y} = \{(\mathbf{x}_i, \mathbf{a}_i, \mathbf{y}_i)\}_{i=1}^n$, classification model $h(X)$, hyperparameter λ.
Output Group-specific threshold $\boldsymbol{\theta} = (\theta_1, \theta_0)$.
Initialize $\boldsymbol{\theta} = (\theta_1, \theta_0) = (0, 0)$.
1. Given a classifier $H(x)$, estimate probability density function $f_{ya}, y, a \in \{0, 1\}$ by maximum likelihood estimation.
while not converge **do**
2. Calculate the optimal step Δ_1 as $\Delta_1 = -\frac{\alpha_1\eta_1 + \lambda\beta_1\epsilon_1}{\alpha_1^2 + \lambda\beta_1^2}$, with $\alpha_1, \beta_1, \eta_1, \epsilon_1$ values shown in (6);
3. Update the threshold: $\theta_1 \leftarrow \theta_1 + \Delta_1$;
4. Calculate the optimal step Δ_0 as $\Delta_0 = -\frac{\alpha_0\eta_0 + \lambda\beta_0\epsilon_0}{\alpha_0^2 + \lambda\beta_0^2}$ with $\alpha_0, \beta_0, \eta_0, \epsilon_0$ values calculated in a similar way as in (6):
5. Update the threshold:$\theta_0 \leftarrow \theta_0 + \Delta_0$.
end while

Taking the derivative of (5) w.r.t. Δ_1 and setting it to 0, we can easily obtain the closed-form solution of Δ_1^τ as

$$\Delta_1^\tau = -\frac{\alpha^\tau \eta^\tau + \lambda \beta^\tau \epsilon^\tau}{(\alpha^\tau)^2 + \lambda(\beta^\tau)^2}. \tag{7}$$

The second step is to fix θ_1 and update θ_0, and this can be achieved in a similar way of updating θ_1. Then we can finalize the alternating optimization as:

$$\theta_0^\tau = \theta_0^{\tau-1} + \Delta_0^\tau, \qquad \theta_1^\tau = \theta_1^{\tau-1} + \Delta_1^\tau. \tag{8}$$

It is notable that in each iteration we derive the optimal update step Δ_a, which eliminates the burden of tuning hyperparameter (such as learning rate) in iterative algorithm. The optimization step is summarized in Algorithm 1. The above algorithm can easily extend to multiple fairness constraints by adding corresponding squared-loss fairness terms to (2).

Upper Bounds on False-Positive/Negative Rate Gap Between Groups

Notations

We start from defining notations. We denote $f_{ya}(x)$ for the estimated parametric probability density function (PDF) of the

distribution of output logit h in the subset $\{Y = y, A = a\}$. Correspondingly, we denote the corresponding cumulative distribution function (CDF) as

$$F_{ya}(x) = \int_{-\infty}^{x} f_{ya}(x)dx.$$

We use $F_{ya}^{-1}(x)$ to denote the inverse of the CDF.

Then, following the definitions given in the main paper, we have

$$\begin{aligned} \text{TP}_a(\theta_a) &= 1 - F_{1a}(\theta_a), & \text{FN}_a(\theta_a) &= F_{1a}(\theta_a), \\ \text{FP}_a(\theta_a) &= 1 - F_{0a}(\theta_a), & \text{TN}_a(\theta_a) &= F_{0a}(\theta_a). \end{aligned} \tag{9}$$

Characterizing the Accuracy Loss Function under Perfect EOp Condition

Before stating the theorem, we illustrate the difference between $\mathcal{L}_{per}(\boldsymbol{\theta})$ used in our paper versus loss function one would use in a population-wise classification problem (without considering group-aware thresholds). That is, one would only consider the loss function on accuracy

$$\bar{\mathcal{L}}_{per}(\theta) = \left(r_1\bar{\text{FN}}(\theta) + r_0\bar{\text{FP}}(\theta)\right)^2, \tag{10}$$

where only one threshold θ (for both groups) needs to be decided, $r_y = (n_{y0} + n_{y1})/N$ is the population ratio of data samples with label y, $\bar{\text{FN}}(\theta), \bar{\text{FP}}(\theta)$ are the population-wise false-negative and false-positive rate. $\bar{\text{FN}}(\theta), \bar{\text{FP}}(\theta)$ are defined in a similar way as in (9) except that we just use the population-wise pdf $\bar{f}_y(x)$ in the integral for label y. (10) will be our benchmark to compare with $\mathcal{L}_{per}(\boldsymbol{\theta})$ used in our paper.

We start from considering the case that we achieve perfect EOp condition, that is

$$\text{TP}_1(\theta_1) = \text{TP}_0(\theta_0), \tag{11}$$

or equivalently

$$\text{FN}_1(\theta_1) = \text{FN}_0(\theta_0).$$

This means that θ_0 and θ_1 satisfies the following condition

$$F_{11}(\theta_1) = F_{10}(\theta_0). \tag{12}$$

Equivalently, we have

$$\theta_0 = F_{10}^{-1}\big(F_{11}(\theta_1)\big). \tag{13}$$

Under any given pair of (θ_0, θ_1) that satisfies (13), recall that the performance error $\mathcal{L}_{per}(\boldsymbol{\theta})$ is defined as

$$\mathcal{L}_{per}(\boldsymbol{\theta}) = \Big(\frac{n_{01}}{N}\text{FP}_1(\theta_1) + \frac{n_{11}}{N}\text{FN}_1(\theta_1) + \frac{n_{00}}{N}\text{FP}_0(\theta_0) + \frac{n_{10}}{N}\text{FN}_0(\theta_0)\Big)^2. \tag{14}$$

From (11), we get

$$\begin{aligned} \frac{n_{11}}{N}\text{FN}_1(\theta_1) + \frac{n_{10}}{N}\text{FN}_0(\theta_0) &= \frac{n_{11}+n_{10}}{N}\text{FN}(\theta_1) \\ &= r_1\text{FN}(\theta_1), \end{aligned}$$

where r_1 denotes, over the entire population (across different groups), proportion of samples with positive labels. In other words, $r_1\text{FN}(\theta_1)$ represents the proportion of data samples (from both groups) with positive label but falsely classified as negative out of the entire dataset.

Next, we look at the other two terms:

$$\frac{n_{01}}{N}\text{FP}_1(\theta_1) + \frac{n_{00}}{N}\text{FP}_0(\theta_0).$$

This sum can be written as

$$\begin{aligned} & \tfrac{n_{01}}{N}\text{FP}_1(\theta_1) + \tfrac{n_{00}}{N}\text{FP}_0(\theta_0) \\ = \; & \tfrac{n_{01}+n_{00}}{N}\text{FP}_1(\theta_1) + \tfrac{n_{00}}{N}\big(\text{FP}_0(\theta_0) - \text{FP}_1(\theta_1)\big) \\ = \; & r_0\text{FP}(\theta_1) + \tfrac{n_{00}}{N}\big(\text{FP}_0(\theta_0) - \text{FP}_1(\theta_1)\big). \end{aligned}$$

We denote $\epsilon_1 = \big(\text{FP}_0(\theta_0) - \text{FP}_1(\theta_1)\big)$. Hence,

$$\mathcal{L}_{per}(\boldsymbol{\theta}) = \mathcal{L}_{per}(\theta_1) = \Big(r_1\text{FN}(\theta_1) + r_0\text{FP}(\theta_1) + \frac{n_{00}}{N}\epsilon_1\Big)^2. \tag{15}$$

Comparing (10) with (15), we can see that, when $\text{FP}_0(\theta_0) > \text{FP}_1(\theta_1)$, the term $\frac{n_{00}}{N}\epsilon_1$ captures the additional accuracy loss due to that we have chosen two different thresholds even though that condition (12) is satisfied. Next, we characterize an upper bound for ϵ_1.

Theorem 1 and Its Proof

Proof. Recall that $\text{FP}_1(\theta_1) = 1 - F_{01}(\theta_1)$ and $\text{FP}_0(\theta_0) = 1 - F_{00}(\theta_0)$. Hence,

$$\begin{aligned} \big|\text{FP}_0(\theta_0) - \text{FP}_1(\theta_1)\big| &= \big|F_{01}(\theta_1) - F_{00}(\theta_0)\big| \\ &\le \big|F_{01}(\theta_1) - F_{01}(\theta_0)\big| \\ &\quad + \big|F_{01}(\theta_0) - F_{00}(\theta_0)\big|. \end{aligned}$$

To bound ϵ, we just need to bound $\big|F_{01}(\theta_1) - F_{01}(\theta_0)\big|$ and $\big|F_{01}(\theta_0) - F_{00}(\theta_0)\big|$.

For the second one, we note that from Assumption 1 that

$$\big|F_{01}(\theta_0) - F_{00}(\theta_0)\big| \le u_0.$$

For the first one, we note that

$$\big|F_{01}(\theta_1) - F_{01}(\theta_0)\big| \le \hat{f}_{01}|\theta_1 - \theta_0|,$$

where $\hat{f}_{01} = \max_x f_{01}(x)$.

Next, we bound $|\theta_1 - \theta_0|$. Note that from (13),

$$\begin{aligned} |\theta_1 - \theta_0| &= \big|F_{10}^{-1}\big(F_{11}(\theta_1)\big) - \theta_1\big| \\ &= \Big|F_{10}^{-1}\big(F_{11}(\theta_1)\big) - F_{10}^{-1}\big(F_{10}(\theta_1)\big)\Big| \\ &\le M_{10}\big|F_{11}(\theta_1) - F_{10}(\theta_1)\big| \\ &\le M_{10}u_1. \end{aligned}$$

□

Theorem 1 provides an upper bound on the difference in the false positive rate between the two groups, for any given pair of (θ_0, θ_1) such that the false negative rates are the same for the two groups (*i.e.,* satisfies the perfect EOp condition). As discussed in Section , this upper bound also characterize the additional accuracy loss due to that we have group-dependent thresholds compared to the case with only one threshold for both groups.

Theorem 2 and its Proof

For predictive equality (PE) condition, we prove a similar result. That is, assuming we achieve perfect PE condition with

$$\text{FP}_1(\theta_1) = \text{FP}_0(\theta_0), \tag{16}$$

or equivalently

$$\text{TN}_1(\theta_1) = \text{TN}_0(\theta_0). \tag{17}$$

This means that θ_0 and θ_1 satisfies the following condition

$$F_{01}(\theta_1) = F_{00}(\theta_0). \tag{18}$$

Equivalently, we have

$$\theta_0 = F_{00}^{-1}\big(F_{01}(\theta_1)\big). \tag{19}$$

Under any given pair of (θ_0, θ_1) that satisfies (19), the performance error $\mathcal{L}_{per}(\boldsymbol{\theta})$ can be written as

$$\begin{aligned}\mathcal{L}_{per}(\boldsymbol{\theta}) &= \Big(\frac{n_{01}}{N}\text{FP}_1(\theta_1) + \frac{n_{11}}{N}\text{FN}_1(\theta_1) \\ &\qquad + \frac{n_{00}}{N}\text{FP}_0(\theta_0) + \frac{n_{10}}{N}\text{FN}_0(\theta_0)\Big)^2 \\ &= \Big(r_1\text{FN}(\theta_1) + r_0\text{FP}(\theta_1) + \frac{n_{10}}{N}\epsilon_2\Big)^2,\end{aligned}$$

where

$$\epsilon_2 = \big(\text{FN}_0(\theta_0) - \text{FN}_1(\theta_1)\big).$$

Similar to Theorem 1, we can provide an upper bound on ϵ_2 under Assumption 1.

Proof. The proof is similar to that of Theorem 1. We provide the main steps and omit details that repeat with the proof of Theorem 1. We have

$$\begin{aligned}\big|\text{FN}_0(\theta_0) - \text{FN}_1(\theta_1)\big| &= \big|F_{11}(\theta_1) - F_{10}(\theta_0)\big| \\ &\le \big|F_{11}(\theta_1) - F_{11}(\theta_0)\big| \\ &\qquad + \big|F_{11}(\theta_0) - F_{10}(\theta_0)\big| \\ &\le \hat{f}_{11}|\theta_1 - \theta_0| + u_1 \\ &\le \hat{f}_{11}M_{00}u_0 + u_1.\end{aligned}$$

□

Theorem 2 provides an upper bound on the difference in the false negative rate between the two groups, for any given pair of (θ_0, θ_1) such that the false positive rates are the same for the two groups (*i.e.,* satisfies the perfect PE condition).

To sum up, Theorem 1 and 2 characterize the upper bound of false positive/negative rate gap between two groups when the false negative/positive rate gap is 0. At the same time, it captures the upper bound of additional accuracy loss due to the two different thresholds for different groups under a perfect fairness (EOp or EP) condition.

Characterizing the Trade-Off between Accuracy and Fairness

In this section, we prove a theorem to characterize the trade-off between accuracy and fairness. That is, we start from the perfect EOp or PE conditions and perturb the solution by a small amount. We then bound the difference in the accuracy loss by comparing the perturbed solution with the original solution that satisfies the perfect fairness conditions.

Perturbed EOp Condition

To start with, let us consider solutions (θ_0, θ_1) that satisfy the perfect EOp condition (13). Under this condition, the optimization problem becomes one dimensional, that is,

$$\theta_1^* = \operatorname*{argmin}_{\theta_1} \mathcal{L}_{per}(\theta_1),$$

where

$$\mathcal{L}_{per}(\theta_1) = \Big(r_1\text{FN}_1(\theta_1) + r_0\text{FP}_1(\theta_1) + \frac{n_{00}}{N}\epsilon_1(\theta_1)\Big)^2$$

and

$$\begin{aligned}\epsilon_1(\theta_1) &= \text{FP}_0(\theta_0) - \text{FP}_1(\theta_1) \\ &= F_{01}(\theta_1) - F_{00}\big(F_{10}^{-1}(F_{11}(\theta_1))\big).\end{aligned}$$

From θ_1^*, we can get the corresponding $\theta_0^* = F_{10}^{-1}(F_{11}(\theta_1^*))$. We further denote this optimal accuracy loss value as

$$L^* = \mathcal{L}_{per}(\theta_1^*).$$

Now with the optimal solution (θ_0^*, θ_1^*), we investigate the changes in $\mathcal{L}_{per}(\theta_1^*)$ when we perturb the perfect EOp condition and allow a small difference. That is, now consider solution $(\tilde{\theta}_0, \tilde{\theta}_1)$ such that

$$\begin{aligned}|\text{FN}_1(\theta_1{}^*) - \text{FN}_1(\tilde{\theta}_1)| &\le \gamma/2, \\ |\text{FN}_0(\theta_0{}^*) - \text{FN}_0(\tilde{\theta}_0)| &\le \gamma/2.\end{aligned} \tag{20}$$

Consequently, the solution $(\tilde{\theta}_0, \tilde{\theta}_1)$ satisfy the following perturbed EOp condition:

$$\big|\text{TP}_1(\tilde{\theta}_1) - \text{TP}_0(\tilde{\theta}_0)\big| = \big|\text{FN}_1(\tilde{\theta}_1) - \text{FN}_0(\tilde{\theta}_0)\big| \le \gamma. \tag{21}$$

Without loss of generality, we assume that (i) the true positive rate of group 1 is higher than that of group 0, and (ii) the above inequality is binding (because if not binding, then we can always choose a smaller γ to make it binding). Thus, we have $\text{TP}_1(\tilde{\theta}_1) - \text{TP}_0(\tilde{\theta}_0) = \gamma$, or equivalently, $\text{FN}_0(\tilde{\theta}_0) - \text{FN}_1(\tilde{\theta}_1) = \gamma$. This gives us

$$\tilde{\theta}_0 = F_{10}^{-1}\big(F_{11}(\tilde{\theta}_1) + \gamma\big). \tag{22}$$

Next, we analyze $\mathcal{L}_{per}(\tilde{\theta}_1)$ by substituting $(\tilde{\theta}_0, \tilde{\theta}_1)$ in (14), which gives us

$$\mathcal{L}_{per}(\tilde{\theta}_1) = \Big(r_1\text{FN}_1(\tilde{\theta}_1) + r_0\text{FP}_1(\tilde{\theta}_1) + \frac{n_{10}}{N}\gamma + \frac{n_{00}}{N}\tilde{\epsilon}_1(\tilde{\theta}_1)\Big)^2,$$

where

$$\begin{aligned}\tilde{\epsilon}_1(\tilde{\theta}_1) &= \text{FP}_0(\tilde{\theta}_0) - \text{FP}_1(\tilde{\theta}_1) \\ &= F_{01}(\tilde{\theta}_1) - F_{00}\big(F_{10}^{-1}\big(F_{11}(\tilde{\theta}_1) + \gamma\big)\big).\end{aligned}$$

We denote the optimal value for this perturbed version as $\tilde{\theta}_1^*$, and its corresponding loss value as

$$\tilde{L}^* = \mathcal{L}_{per}(\tilde{\theta}_1^*).$$

Furthermore, from (20), we have

$$|\text{FN}_1(\theta_1^*) - \text{FN}_1(\tilde{\theta}_1^*)| = |F_{11}(\theta_1^*) - F_{11}(\tilde{\theta}_1^*)| \le \gamma/2. \tag{23}$$

Under Assumption 1, we have

$$\begin{aligned}|\theta_1^* - \tilde{\theta}_1^*| &= \big|F_{11}^{-1}(F_{11}(\theta_1^*)) - F_{11}^{-1}(F_{11}(\tilde{\theta}_1^*))\big| \\ &\le M_{11}\big|F_{11}(\theta_1^*) - F_{11}(\tilde{\theta}_1^*)\big| \\ &= M_{11}\gamma/2.\end{aligned}$$

Theorem 3 and Its Proof

We are ready to compare $\mathcal{L}_{per}(\theta_1^*)$ and $\mathcal{L}_{per}(\tilde{\theta}_1^*)$. The latter loss should be no larger than the former since we relaxed the perfect EOp condition (constraint) in the optimization, *i.e.,* $L^* \geq \tilde{L}^*$.

Proof. We have that

$$\begin{aligned}
& \mathcal{L}_{per}(\theta_1^*) - \mathcal{L}_{per}(\tilde{\theta}_1^*) \\
\leq\ & 2L^* \Big| r_1 \text{FN}_1(\theta_1^*) + r_0 \text{FP}_1(\theta_1^*) + \frac{n_{00}}{N}\epsilon_1(\theta_1^*) \\
& - \Big(r_1 \text{FN}_1(\tilde{\theta}_1^*) + r_0 \text{FP}_1(\tilde{\theta}_1^*) + \frac{n_{10}}{N}\gamma + \frac{n_{00}}{N}\tilde{\epsilon}_1(\tilde{\theta}_1^*) \Big) \Big| \\
\leq\ & 2L^* \big(r_1\gamma/2 + r_0 |\text{FP}_1(\theta_1^*) - \text{FP}_1(\tilde{\theta}_1^*)| \\
& + \frac{n_{00}}{N} \big| \epsilon_1(\theta_1^*) - \tilde{\epsilon}_1(\tilde{\theta}_1^*) \big| + \frac{n_{10}}{N}\gamma \big),
\end{aligned}$$

where we further have that

$$\begin{aligned}
|\text{FP}_1(\theta_1^*) - \text{FP}_1(\tilde{\theta}_1^*)| &= |F_{01}(\theta_1^*) - F_{01}(\tilde{\theta}_1^*)| \\
&\leq \hat{f}_{01} |\theta_1^* - \tilde{\theta}_1^*| \\
&\leq \hat{f}_{01} M_{11} \gamma/2,
\end{aligned}$$

and

$$\begin{aligned}
\big| \epsilon_1(\theta_1^*) - \tilde{\epsilon}_1(\tilde{\theta}_1^*) \big| &\leq \big| \epsilon_1(\theta_1^*) - \tilde{\epsilon}_1(\theta_1^*) \big| \\
&\quad + \big| \tilde{\epsilon}_1(\theta_1^*) - \tilde{\epsilon}_1(\tilde{\theta}_1^*) \big| \\
&\leq \big| F_{00}(F_{10}^{-1}(F_{11}(\theta_1^*))) \\
&\quad - F_{00}(F_{10}^{-1}(F_{11}(\theta_1^*) + \gamma)) \big| \\
&\quad + \hat{\epsilon}_1' M_{11} \gamma/2 \\
&= (\hat{f}_{00} M_{10} + \hat{\epsilon}_1' M_{11}/2)\gamma.
\end{aligned}$$

Here, $\hat{\epsilon}_1' = \max \tilde{\epsilon}_1'$ is the maximum of the derivative of $\tilde{\epsilon}_1$. Combining all the terms in front of γ gives us the desired upper bound. □

Theorem 3 quantifies the decrease in accuracy loss (*i.e.,* the improvement in accuracy) when we allow a gap of true positive rates between two groups (*i.e.,* relaxation from the perfect EOp condition).

Repeating the analysis for the perturbed PE condition, we can obtain a similar bound for the changes in the accuracy loss function. We omit the details here in the interest of space.

Convergence Analysis of GSTAR

GSTAR as Nonlinear Least Squares Problem

Our objective function and the optimization solution algorithm belong to the family of **Gauss-Newton algorithm** to solve Nonlinear Least Squares Problem (NLSP). To specify, NLSP is to solve

$$\min_{\boldsymbol{\theta}} ||r(\boldsymbol{\theta})||_2^2,$$

where the decision variables, $\boldsymbol{\theta}$, is an n-dimensional real vector and the objective function r is an m-dimensional real vector function of $\boldsymbol{\theta}$. Connecting to our setting and using two groups as an example, our decision variables is the two-dimensional vector $\boldsymbol{\theta} = (\theta_0, \theta_1)$ for group 0 and group 1, and our objective function is the following 2-dimensional real vector function:

$$r(\boldsymbol{\theta}) = \big(r_1(\boldsymbol{\theta}), r_2(\boldsymbol{\theta}) \big)$$

with

$$\begin{aligned}
r_1(\boldsymbol{\theta}) = r_1(\theta_0, \theta_1) &= \frac{n_{01}}{N}\text{FP}_1(\theta_1) + \frac{n_{11}}{N}\text{FN}_1(\theta_1) \\
&\quad + \frac{n_{00}}{N}\text{FP}_0(\theta_0) + \frac{n_{10}}{N}\text{FN}_0(\theta_0), \\
r_2(\boldsymbol{\theta}) = r_2(\theta_0, \theta_1) &= \sqrt{\lambda}\,(\text{TP}_1(\theta_1) - \text{TP}_0(\theta_0))
\end{aligned}$$

when taking the EOp constraint. The L_2 norm $||r(\boldsymbol{\theta})||_2^2 = r_1(\boldsymbol{\theta})^2 + r_2(\boldsymbol{\theta})^2$ recovers the objective function in Equation (2) in the main paper.

A classic family of algorithms to solve NLSP is the Gauss-Newton Method. The main idea is to convert the nonlinear optimization problem to a linear least square problem using Taylor expansion. That is, the parameter values are calculated in an iterative fashion with

$$\theta_j \approx \theta_j^{k+1} = \theta_j^k + \Delta_j,$$

in the k-th iteration number, with the vector of increments $\Delta = \{\Delta_j\} = \{\theta_j^{k+1} - \theta_j^k\}$ (also known as the shift vector). We linearize each component in the f function to a first-order Taylor polynomial expansion as

$$r_i(\boldsymbol{\theta}) \approx r_i(\boldsymbol{\theta}^k) + \sum_j \frac{\partial r_i(\boldsymbol{\theta}^k)}{\partial \theta_j} \Delta_j \tag{24}$$

with $\boldsymbol{\theta}^k = (\theta_0^k, \theta_1^k)$. Plugging this linearized equation into the objective function, we get the usual least square problem. Then, the optimal solution can be obtained as

$$\Delta = -(J^T J)^{-1} J^T f(\boldsymbol{\theta}^k), \tag{25}$$

where $J = \{J_{ij}\}$ with $J_{ij} = \{\frac{\partial r_i(\boldsymbol{\theta})}{\partial \theta_j}\}$ is the Jacobian. Note that in the GSTAR algorithm, we ignore the terms for $j \neq i$ in the Taylor expansion (24). Thus, in calculating $J^T J$, we only kept the diagonal terms

$$\left(\frac{\partial r_1(\boldsymbol{\theta})}{\partial \theta_j} \right)^2 + \left(\frac{\partial r_2(\boldsymbol{\theta})}{\partial \theta_j} \right)^2$$

for $j = 0, 1$. Plugging in the form of r_1 and r_2 as specified above, we achieve the solution provided in (7).

Convergence Property for Gauss-Newton Algorithm

There is a long history of studying the convergence property of the Gauss-Newton algorithm, *e.g.,* see (Gratton, Lawless, and Nichols 2007). The convergence of the algorithm is generally not guaranteed, *e.g.,* if the initial solution is far from the true optimal or $J^T J$ is ill-conditioned. In other words, the convergence of the algorithm heavily depends on the density estimation $f(\cdot)$. We state the following sufficient conditions from (Gratton, Lawless, and Nichols 2007) to guarantee the convergence of the algorithm. The following assumptions are made in order to establish the theory.

- A1. There exists θ^* such that $J^T(\theta^*)r(\theta^*) = 0$;
- A2. The Jacobian at θ^* has full rank.

We state Theorem 4 from (Gratton, Lawless, and Nichols 2007) on the sufficient conditions for convergence.

Theorem 4. *Assume that the estimated density function $f(\cdot)$ satisfy assumptions A1 and A2. Further, $f(\cdot)$ satisfies that*

$$||Q(\theta^k)(J^TJ)^{-1}(\theta^k)||_2 \leq \eta$$

for some constant $\eta \in [0, 1)$ for each iteration k, where $Q(\theta)$ denotes the second order terms $\sum_i r_i(\theta)\nabla^2 r_i(\theta)$. Then as long as the initial solution is sufficiently close to the true optimal with $||\theta^0 - \theta^*||_2 \leq \epsilon$, the sequence of Gauss-Newton iterates $\{\theta^k\}$ converges to θ^*.

Protection against Divergence

It is known that for general function $f(\cdot)$ such as estimates from a neural network, the above sufficient conditions that guarantee convergence do not necessarily hold. As a result, protection against divergence is essential. In our numerical experiments, we adopt a commonly used, simple protection, the shift-cutting method. That is, we to reduce the length of the shift vector Δ by a fraction η. In other words, the update becomes

$$\theta_j^{k+1} = \theta_j^k + \eta\Delta_j.$$

Near Optimality of GSTAR

Here, we show regarding on how the accuracy of h affects the accuracy of $\hat{Y}$. Following the proof of Theorem 5.6 of Hardt *et al.* (2016), we provide the following near optimality results for our method.

Before we prove the theorem, we first state the results from Lemma 5.5 proved in Hardt et al. (2016), which will be used in our proof.

Lemma 5 (Restatement of Lemma 5.5 in Hardt et al. (2016))**.** *Let $R, R' \in [0, 1]$ be two random variables in the same probability space as A and Y. Then, for any point p on a conditional ROC curve of R, there is a point q on the corresponding ROC curve of R' achieving the same threshold such that*

$$\begin{aligned} ||p - q||_2 &\leq \sqrt{2}d_K(R, R'), \\ d_K(R, R') = \max_{a,y} \sup_t &|Pr(R > t|A = a, Y = y) \\ &- Pr(R' > t|A = a, Y = y)|. \end{aligned} \tag{26}$$

Proof. Similar to Hardt et al. (2016), we focus on proving this theorem for equalized odds. The case of equal opportunity is analogous. The optimal classifier Y^* corresponds to a point p^*. Under the equalized odds condition, our algorithm essentially finds the intersection point, q, of the two conditional ROC curves of $\hat{R}_h$ for $a = 0$ and $a = 1$. Then directly applying the above lemma, we get that

$$||p^* - q||_2 \leq \sqrt{2}d_K(R, R').$$

The rest follows the same argument as in Theorem 5.6 of Hardt et al. (2016). That is, by assumption on the loss function, there is a vector v with $||v||_2 \leq \sqrt{2}$ such that $\mathbb{E}[\ell(\hat{Y}_h, Y)] = \langle v, q\rangle$ and $\mathbb{E}[\ell(Y^*, Y)] = \langle v, p^*\rangle$. Applying Cauchy-Schwarz, we get

$$\begin{aligned} \mathbb{E}[\ell(\hat{Y}_h, Y)] - \mathbb{E}[\ell(Y^*, Y)] &= \langle v, q - p^*\rangle \\ &\leq ||v||_2 \cdot ||p^* - q||_2 \\ &\leq 2d_K(R, R'). \end{aligned}$$

□

Remark. *In Hardt et al. (2016), the point q from their algorithm under equalized odds condition is the intersection point between two line segments, not the two ROC curves as in our paper. That is, assume without loss of generality that the first coordinate of q_1 (for group $a = 1$) is greater than the first coordinate of q_0 (for group $a = 0$) on the ROC curve plane; and that all points p^*, q_0, q_1 lie above the main diagonal. Then $q \in L_0 \cap L_1$ from their algorithm, where L_0 is the line segment between q_0 and the point $(1, 1)$, and L_1 is the line segment between the point $(0, 0)$ and q_1. As a result, in proving their Theorem 5.6, they need to show that q from this construction satisfies $||p^* - q||_2 \leq 2d_K(R, R')$. However, because the point q from our algorithm lies on the ROC curve, we can directly apply the results from the lemma. This difference is further illustrated in figure 5 below, where the purple pentagram corresponds to q found by our algorithm, and the green cross corresponds to q from their algorithm. The figure shows the intersection points found from our algorithm versus Hardt et al.*

Moreover, the requirement for achieving the near optimality in our method (our Theorem 5) and in Hardt et al. (their Theorem 5.6) is the same. That is, the closeness between the conditional densities is required, not just the conditional probability estimates.

To specify, the closeness requirement in Hardt et al. based on conditional Kolmogorov distance is shown in Equation (26)*, where $R \in [0, 1]$ and $R' \in [0, 1]$ are real-valued scores,* i.e., *two regressors. Note that the distance is taking* sup *over all $t \in [0, 1]$, so this condition requires the entire conditional density curves from R and R' to be close, not just close at some given threshold t.*

Next, the near optimality of Hardt et al. (their Theorem 5.6) shows:

$$\mathbb{E}[\ell(\hat{Y}_h, Y)] \leq \mathbb{E}[\ell(Y^*, Y)] + 2\sqrt{2}d_K(\hat{R}, R^*),$$

where $R^* \in [0, 1]$ is the Bayes optimal regressor and $\hat{R} \in [0, 1]$ is a regressor whose density is estimated. In fact, the distribution function of their corresponds to the score function $\sigma(h)$ in our paper, where h is the logit and $\sigma(\cdot)$ is the softmax function.

Seeing this connection, we stress that the closeness requirement in our result is the same as in Hardt et al., and that the near optimality of our algorithm follows:

$$\mathbb{E}[\ell(\hat{Y}_h, Y)] \leq \mathbb{E}[\ell(Y^*, Y)] + 2d_K(\hat{R}, R^*),$$

where $R^* \in [0, 1]$ is the Bayes optimal regressor as given in Hardt et al., and $\hat{R} \in [0, 1]$ comes from our estimated density, i.e., *, the distribution of $\hat{R}_h$ comes from by applying softmax function $\sigma(\cdot)$ on logit h.*

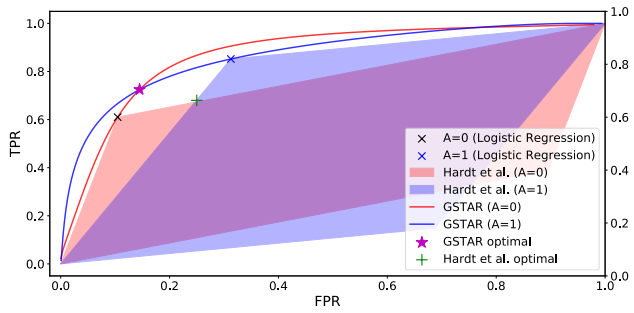


Figure 5: Comparison of optimal point of Hardt *et al.*and GSTAR. Given the ROC curve of each protected group, ours (magenta star) achieves better optimum than that of Hardt *et al.*(green cross), as ours has higher TPR and lower FPR.

Experimental Details

Comparing Methods

We compared our method with multiple state-of-the-art methods to verify our work. The details about the comparing methods are as below:

- **Learning fair representations for kernel models** (abbreviated as FGP) (Tan et al. 2020): a pre-processing method to learn representation focusing on kernel-based models. The fair model that satisfies certain fairness criterion is obtained by Bayesian learning from fair Gaussian process (FGP) prior.
- **Fairness confusion tensor** (abbreviated as FACT) (Kim, Chen, and Talwalkar 2020): a post-processing model that minimize the least-squares accuracy-fairness optimality problem based on confusion tensor.
- **Adversarial de-biasing** (abbreviated as AdvDeb) (Zhang, Lemoine, and Mitchell 2018): an in-processing model that mitigates the conflicting gradient directions in utility and fairness objectives by projecting one gradient to another to remove the opposite direction.
- **Calibrated equal odds post-processing** (abbreviated as CEOPost) (Pleiss et al. 2017): a post-processing method that minimizes the disparity in the predicted probability to the preferred class among different sensitive groups, while maintaining the calibration condition in a relaxed condition.
- **Equality of opportunity in supervised learning** (abbreviated as Eq.Odds) (Hardt, Price, and Srebro 2016): a post-processing method that learns the threshold to yield the equalized odds/opportunity between different demographic by exploring the intersection of achievable true positive rates and false positive rates.
- **Learning adversarially fair and transferable representations** (abbreviated as LAFTR) (Madras et al. 2018): a fair representation learning model that adopts fairness metrics as the adversarial objectives and analyze the balance between utility and fairness.
- **Baseline**: For CelebA dataset, we use ResNet50 (He et al. 2016) as a reference because we input second last layer (2048 features) of ResNet to all methods. For other tabular datasets, logistic regression is used as all other methods except for FGP and LAFTR are based on logistic regression.

Evaluation Metrics

In the experiments, we evaluate the methods on four fairness and two performance measures. Four fairness metrics are as below:

- **Equal Opportunity** (abbreviated as EOp) (Hardt, Price, and Srebro 2016) : This measures absolute difference of favorable prediction given positive label.

$$|P(\hat{Y}=1|Y=1,A=1)-P(\hat{Y}=1|Y=1,A=0)|.$$

- **Equalized Odds** (abbreviated as EOd) (Hardt, Price, and Srebro 2016) : This measures the difference between the probability given the true labels.

$$|P(\hat{Y}=1|Y=1,A=1)-P(\hat{Y}=1|Y=1,A=0)|+$$
$$|P(\hat{Y}=1|Y=0,A=1)-P(\hat{Y}=1|Y=0,A=0)|.$$

- **Balanced Accuracy Difference** (abbreviated as BD) : This measures difference between balanced accuracy between the groups.

$$\Big|P(\hat{Y}=1|Y=1,A=1)+P(\hat{Y}=0|Y=0,A=1)\Big|$$
$$-\Big|P(\hat{Y}=1|Y=1,A=0)+P(\hat{Y}=0|Y=0,A=0)\Big|.$$

  If BD and EOd has the same value, it indicates that both TPR and TNR are higher in a certain sensitive group. However, if the gap between the two terms is large, we can interpret as the classifier is more biased as a group with higher TPR has lower TNR. This is more unfair as a sample from the privileged group is more likely to be falsely and correctly predicted as positive output. EOp is a partial measure of EOd as it only measures the difference from a favorable class.
- **Absolute (1 - Disparate Impact)** (abbreviated as 1-DIMP) (Barocas and Selbst 2016) : This measures ratio of the probability of the favorable prediction given a protected group.

$$\left|1-\frac{P(\hat{Y}=1|A=1)}{P(\hat{Y}=1|A=0)}\right|.$$

We evaluate performance of the methods with two metrics.

- **Balanced Accuracy** (abbreviated as BA) : This measures average between true positive rate and true negative rate. Compared to the traditional accuracy, this measure effectively shows the whether the classifier is focusing on the performance of a certain class when the dataset is unbalanced.

$$\frac{1}{2}\Big(P(\hat{Y}=1|Y=1)+P(\hat{Y}=0|Y=0)\Big).$$

- **Accuracy** (abbreviated as ACC) : This measures traditional classification accuracy of the method.

Experimental Setup

For experimental setup, all comparing methods apply EOd as the fair constraint if fairness constraint is selectable, thus we compare them via EOd in Figure 2 in the main paper. Both the Pareto frontier from GSTAR and FACT are derived based on EOd constraint for a fair comparison. We follow the setup in Section G.3 of the FACT (Kim, Chen, and Talwalkar 2020) to report their method, which does not require λ.

For GSTAR, we estimate f_{ya} and optimize θ_a from the training data, and report evaluation results (with the θ_a learned from training data) on the testing data. We use the same λ for multiple fairness constraints for simplicity, but λ can be introduced individually. Our method is optimized with λ in the range of $[10^{-1}, 10^4]$ with alternating optimization method.

To find estimated distribution f_{ya}, we consider gamma, Student's t, and normal distribution as the candidates for the experiments reported in the main paper, and select the one that has the maximum likelihood with the output distribution. Without loss of generality, this can be generalized non-parametric density estimation such as kernel density estimation to fit more complicated distribution. More experiments with complicated distribution estimation is in Section in the supplementary.

Figure 2 illustrates Pareto frontiers with 5 points of different λ values in $[10^{-2}, 10^7]$ with equal logspace. To visualize the plots, we sweep hyperparameters (e.g, weights for each term in the objective function) for comparing methods. Figure 3 takes λ or hyperparameter values from the upper-right point of the Pareto frontiers in Figure 2, which indicates the best trade-off for each method. Figure 3 paper presents the 5 runs with the setup chosen based on the Pareto frontier to show the consistency of the performance of each model.

All experiments are implemented with Pytorch framework on i9-9960X CPU and a Quadro RTX 6000 GPU.

Dataset Description

We evaluate the methods on four fairness datasets. The goal for all datasets is binary classification on binary sensitive feature. The details of the datasets are as below:

- **CelebA image dataset**[1] (Liu et al. 2015): The data consists of 202,599 face images in diverse demographics. The images are annotated with 40 attributes (face shape, skin tone, smiling, etc.). Similar to Quadrianto *et al.* (Quadrianto, Sharmanska, and Thomas 2019), the goal is to predict whether a person in the image is attractive or not. The feature *sex* is used as the sensitive feature.
- **Adult** dataset from the UCI repository (Kohavi 1996) contains 48,842 instances described by 14 features (workclass, age, education, sex, race, *etc*) with the goal of the income prediction whether a person's income exceeds 50K USD per year. The feature *sex* is used as the sensitive feature.
- **COMPAS**[2](Correctional Offender Management Profiling for Alternative Sanctions) dataset includes 6,167 samples described by 401 features with the target of recidivism prediction with the label showing if each person gets rearrested within two years. The feature *race* is used as the sensitive feature for this dataset.
- **German** credit dataset from the UCI repository (Dua and Graff 2019) contains 1,000 samples described by 20 features. The goal is to predict the credit risks. The feature *sex* is used as the sensitive feature.

All data is split as 70% for training and 30% for testing.

[1]http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

[2]https://github.com/propublica/compas-analysis

CelebA				
Model	GSTAR	FGP	FACT	CEOPost
Time	0.287	-	0.067	0.077
Model	DIR	Eq.Odds	LAFTR	AdvDeb
Time	183.20	0.062	107.04(min)	303.15
Adult				
Model	GSTAR	FGP	FACT	CEOPost
Time	0.29	51.28	0.055	25.61
Model	DIR	Eq.Odds	LAFTR	AdvDeb
Time	168	0.037	53.04(min)	102.00
Compas				
Model	GSTAR	FGP	FACT	CEOPost
Time	0.292	43.74	0.035	8.3
Model	DIR	Eq.Odds	LAFTR	AdvDeb
Time	123.20	0.034	57.04(min)	15.45
German				
Model	GSTAR	FGP	FACT	CEOPost
Time	0.271	7.08	0.0257	2.64
Model	DIR	Eq.Odds	LAFTR	AdvDeb
Time	1.68	0.034	56.51(min)	2.17

Table 1: Computational time (in seconds) for all comparing fairness methods for each dataset.

Computational Cost

In Table 1, we describe the computational time for each method on each dataset. By introducing estimated PDF functions for post-processing, we outperform other methods except Eq.Odds (Hardt, Price, and Srebro 2016) and FACT (Kim, Chen, and Talwalkar 2020). As they both only utilize the entries of the confusion matrix to find optimal mixing rate in their methods, they have less computation than ours. However, as we discussed in the main paper, we explore better feasible region than theirs by group-specific thresholding that results better in both fairness and performance by sacrificing little efficiency, yet outperforms most of the other works.

Auxiliary Experiments

GSTAR with Single Threshold

We conduct experiments on COMPAS dataset to evaluate GSTAR with a single adaptive thresohld. Figure 7 presents the trend of fairness-accuracy trade-off of two versions of GSTAR based on λ values. Comparing with the baseline ($\theta = 0$), we observe that even with a single threshold in

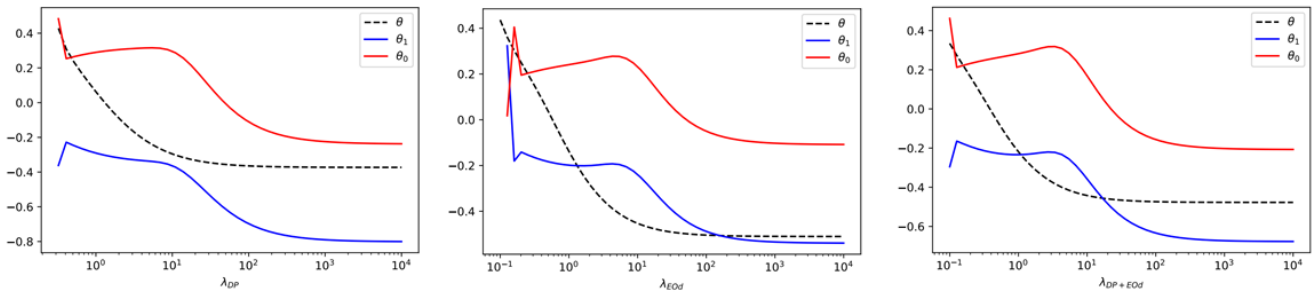


Figure 6: Trend of converged θ values based on the variation of weight λ. Dashed line indicates single threshold version and θ_a indicates threshold for a group.

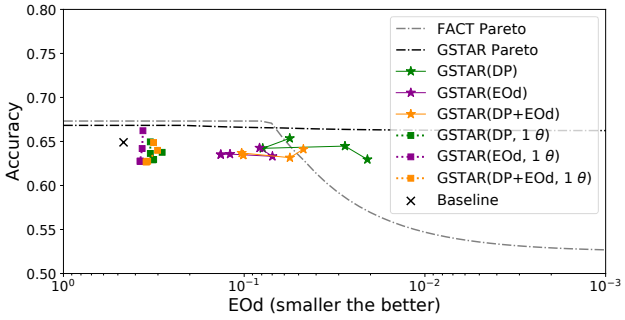


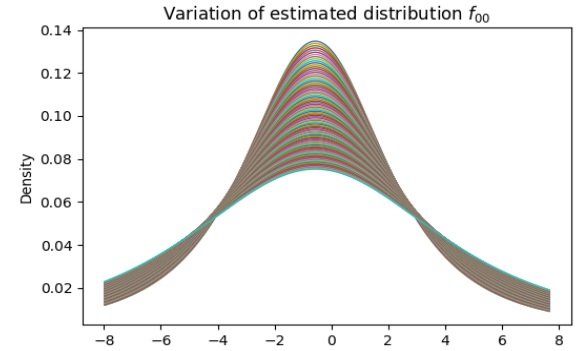
Figure 7: Comparison of single threshold (squares, 1θ) and group-aware threshold method (stars) on Pareto frontier. The result suggests group-aware threshold greatly improve fairness with comparable accuracy.

GSTAR (1θ in the legend), the adaptive threshold helps to improve the fairness with comparable accuracy. However the improvement is not as significant as that of the group-wise version because it is impossible to achieve perfect fairness with a single threshold as the intersection of f_{1a} and f_{0a} differs by a. Figure 6 shows the trend of learned θ based on λ values. We see a single threshold version (black) lies between two thresholds of group-aware GSTAR in most cases. This implies that the single threshold converges to some point that gives up some of the fairness.

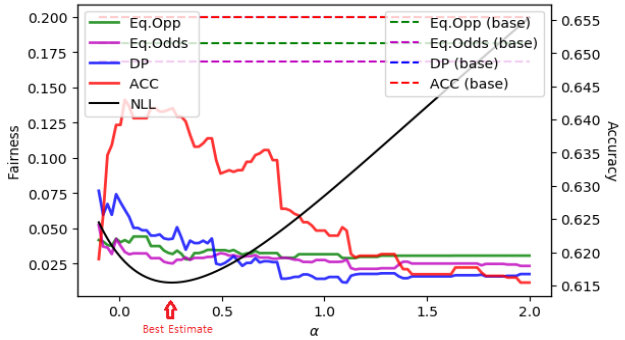
Quality of Estimated Distribution

The performance of GSTAR relies on the quality of estimated distribution. For the benchmark datasets, we empirically found that the distribution of logits resembles some parametric distributions. Thus, we estimate the distribution with generally used parametric distributions such as Student's t-distribution by measuring the negative log-likelihood (NLL) in the training data. Note that GSTAR can be extended to a wide range of other distributions, even non-parametric distributions.

For further analysis, we add new experiments by sweeping the parameters of parametric distribution to see the effect of the estimation quality. In COMPAS dataset, the best estimate



(a) Variation of estimated distribution of f by the noise factor α.



(b) The influence of the noise factor α and NLL of corresponding estimated distribution to the performance and fairness of GSTAR.

Figure 8: Variation of estimated distributions by the noise α and its impact on the performance of GSTAR.

(*i.e.,* smallest NLL) of group $(y = 0, a = 0)$ with Student's t-distribution has parameters of df = 2.235, loc = -0.567, scale = 0.756 based on scipy package. To generate variations as in Figure 8(a) of distributions with varying estimation qualities, we add noise $\alpha \in [-0.1, 100]$ to the scale of this distribution.

In figure 8(b), we illustrate the trend of NLL (black), fairness violation (the lower the better), and accuracy (the higher the better) with varying noise (α, x-axis). The color of lines follows the main paper. Dashed lines indicate the quantity of baseline model ($\theta = 0$). From this, we observed that the

	ACC (train)	DP (train)	EOd (train)	ACC (test)	DP (test)	EOd (test)
GSTAR (DP)	0.679	0.001	0.089	0.639	0.017	0.018
GSTAR (EOd)	0.714	0.071	0.030	0.0643	0.032	0.034
GSTAR (DP + EOd)	0.705	0.050	0.030	0.641	0.027	0.025

Table 2: Comparison of fairness and performance measure in training and testing set with different fairness constraints.

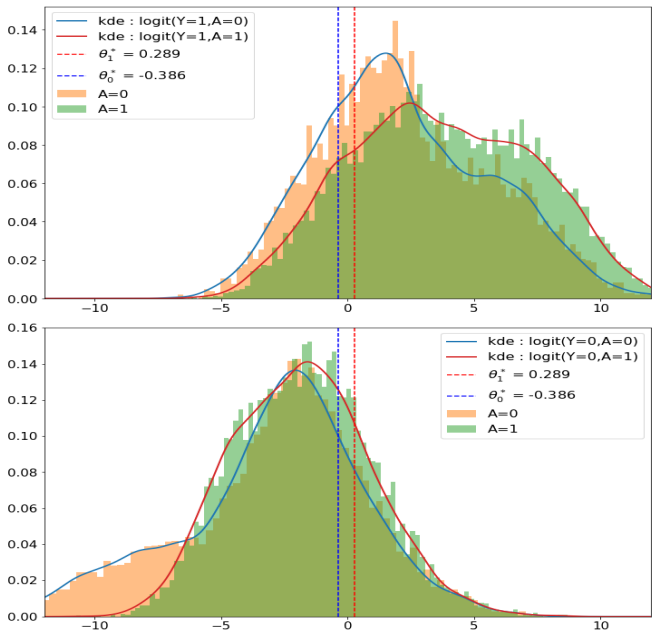


Figure 9: Distribution of synthetic dataset and its kernel density estimation.

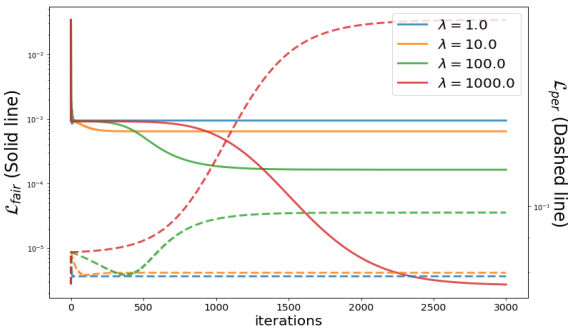


Figure 10: Convergence analysis of GSTAR on COMPAS dataset with different λ values.

accuracy is the most sensitive to the change of estimation quality, while fairness is relatively stable.

However, for our experiments, we assume the estimation is reliable and the guarantee on the estimation reliability is beyond our focus of this paper.

Interpretation of Results on COMPAS

In COMPAS in Figure 3(c), we observe improvements in total fairness violation with multiple fairness constraints employed. We deduce this could happen due to: 1) generalization of the estimated distribution from training data to testing data; 2) difference in the training and testing data distributions. For the training set, we achieve better fairness violation on the model with a single constraint, compared to the multi-constrained version or other single-constrained versions. In the training set of COMPAS data, we have the results as in the Table 2.

Empirical Result on Convergence of GSTAR

In Theorem 4, we showed that convergence proof given the conditions that 1) GSTAR satisfies Jacobian conditioning; and 2) initial solution θ^0 is sufficiently close to θ^*. To verify the convergence in practice, we illustrate the change of $\mathcal{L}_{fair}$ and $\mathcal{L}_{per}$ values by epochs as in Figure 10 in COMPAS dataset. We found that it takes longer to converge as λ grows. However, regardless of λ, we observed that GSTAR successfully converged. Considering the difference between initial and converged values of $\mathcal{L}_{fair}$ and $\mathcal{L}_{per}$, it seems that θ^* differs from θ^0 by a larger amount as λ grows, and this leads to longer convergence time (epochs).

Complicated Distribution Estimation with Kernel Density Estimation

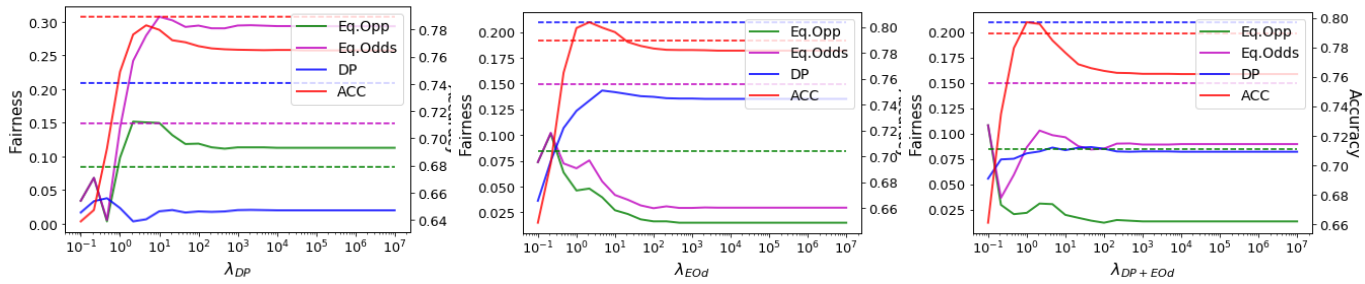
We can generalize our density estimation to non-parametric by kernel density estimation (KDE) method. Given the logit distribution $h(X)$, we build a histogram with B bins. Denote T_b as the mean logit value of b-th bin and w_b as normalized weight indicates how many samples belong to b-th bin, where $b \in \{1, \cdots, B\}$ and $\sum_b w_b = 1$. Then our kernel density estimator of distribution $h(X)$ is

$$f(x) = \sum_b w_b K(x - T_b), \tag{27}$$

where K is kernel function and we employ normal distribution with standard deviation as 0.5. As non-parametric density estimation of $h(x)$ can be expressed as linear combination of parametric distributions, we can easily apply the optimization step demonstrated in the Section .

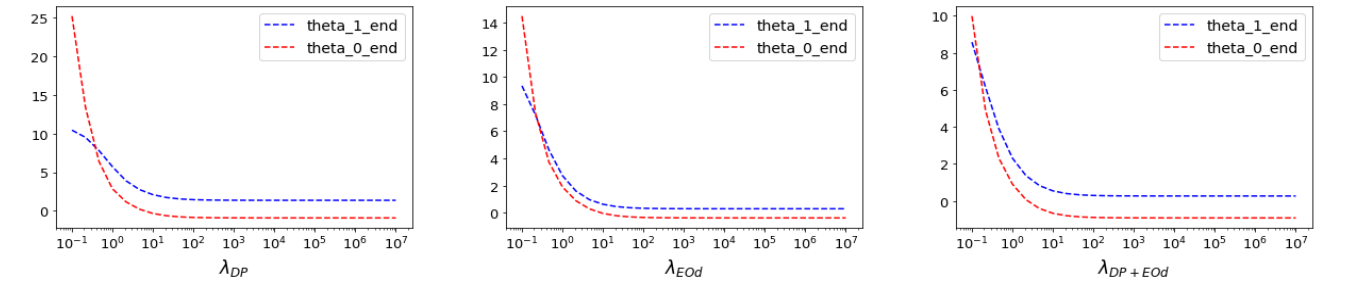
To validate the KDE method for GSTAR, we generate synthetic data that each logit distribution h_{ya} consist of mixture of three gaussian distributions with additional standard normal noise. Specifically, each distribution is configured with mean μ, variation σ^2, and weight w and number of samples n as in Table 3. For example, we generate the samples from a group $(Y = 0, A = 0)$ by sampling n_{00} samples x on h_{00} and add noise $\mathcal{N}(0, 1^2)$ as below:

$$h_{00} = \sum_{i \in \{0,1,2\}} w_{00}^{(i)} \cdot \mathcal{N}(\mu_{00}^{(i)}, {\sigma_{00}^{(i)}}^2),$$



(a) $\mathcal{L}_{fair} = \mathcal{L}_{fair}^{DP}$ (b) $\mathcal{L}_{fair} = \mathcal{L}_{fair}^{EOd}$ (c) $\mathcal{L}_{fair} = \mathcal{L}_{fair}^{DP+EOd}$

Figure 11: Trend of performance and fairness measure by the change of λ values. Color of the lines indicates the measure of performance and fairness as in the legend. Solid lines indicate GSTAR results and dotted lines indicate baseline ($\theta = (0, 0)$) respectively. It is lower the better for fairness and higher the better for accuracy.



(a) $\mathcal{L}_{fair} = \mathcal{L}_{fair}^{DP}$ (b) $\mathcal{L}_{fair} = \mathcal{L}_{fair}^{EOd}$ (c) $\mathcal{L}_{fair} = \mathcal{L}_{fair}^{DP+EOd}$

Figure 12: Trend of converged group-aware threshold θ achieved by GSTAR.

$$l^{(k)} \sim h_{00}, \quad \epsilon^{(k)} \sim \mathcal{N}(0, 1^2),$$
$$x^{(k)} := l^{(k)} + \epsilon^{(k)}, \quad k \in \{1, \cdots, n_{00}\}$$

where i is the index of gaussian distributions in Table 3 and k is the index of sampling instance.

	μ	σ^2	w	n
h_{00}	[-7.0, -2.0, 1.1]	[3.0, 1.5, 2.0]	[0.3, 0.5, 0.2]	5000
h_{01}	[-4.5, -1.2, 1.2]	[1.2, 1.5, 2.0]	[0.3, 0.5, 0.2]	10000
h_{10}	[-1.8, 1.5, 6.0]	[1.2, 1.3, 2.0]	[0.2, 0.5, 0.3]	15000
h_{11}	[-1.1, 2.3, 7.0]	[1.2, 1.5, 2.0]	[0.2, 0.4, 0.4]	10000

Table 3: Configuration of each synthetic data distribution h_{ya}.

Figure 9 illustrates histograms of logit h distributions of synthetic data and their KDE results in colored lines. The top plot is about positive samples *i.e.,* h_{11} and h_{10}, and the bottom plot is about positive samples *i.e.,* h_{00} and h_{01} respectively. We could observe that KDE accurately estimated the density function h that cannot be fitted with parametric distribution.

Moreover, we conduct experiments to validate GSTAR can achieve the proposed goal. Given 4 probability distributions and number of samples for each group as in Table 3, we divide the dataset into training (70%), validation (15%), and testing (15%) set. We train GSTAR on training set and find the best $\boldsymbol{\theta}$ by selecting one that has minimum validation loss and report the result on testing set.

In Figure 11 and 12, we quantitatively evaluate GSTAR with KDE method with different λ values on fairness constraint $\mathcal{L}_{fair}$. In Figure 11, color of the lines are performance and fairness measure as described in the legend. Dotted lines indicate baseline ($\boldsymbol{\theta} = (0, 0)$) and solid lines indicate the measures of GSTAR. Note that GSTAR improve target fairness significantly with small lose of accuracy. In DP+EOd constraint, we even achieve almost perfect equal opportunity *i.e.,* Eq. Opp ≈ 0 with high enough λ values.

References

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

Dua, D.; and Graff, C. 2019. UCI Machine Learning Repository. *University of California, Irvine, School of Information and Computer Sciences*.

Gratton, S.; Lawless, A. S.; and Nichols, N. K. 2007. Approximate Gauss–Newton methods for nonlinear least squares problems. *SIAM*, 18(1): 106–132.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *NIPS*, 3315–3323.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Kim, J. S.; Chen, J.; and Talwalkar, A. 2020. Model-Agnostic Characterization of Fairness Trade-offs. *arXiv preprint arXiv:2004.03424*.

Kohavi, R. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, volume 96, 202–207.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *ICCV*, 3730–3738.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On fairness and calibration. In *NIPS*, 5680–5689.

Quadrianto, N.; Sharmanska, V.; and Thomas, O. 2019. Discovering fair representations in the data domain. In *CVPR*, 8227–8236.

Tan, Z.; Yeom, S.; Fredrikson, M.; and Talwalkar, A. 2020. Learning fair representations for kernel models. In *AISTATS*, 155–166.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *AIES*, 335–340.