Unified Fairness from Data to Learning Algorithm

Yanfu Zhang, Lei Luo, Heng Huang
Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, United States
yaz91@pitt.edu, zzdxpyyll@gmail.com, henghuanghh@gmail.com

***Abstract*—In classification problems, individual fairness prevents discrimination against individuals based on protected attributes. Fairness-aware methods usually consist of two stages, first determining a fair metric concerning the similarity between different instances and then learning the fairness-aware model. However, existing works usually consider these two stages separately and only focus on improving the individual stage. Moreover, the choice of fair metric is heavily dependent on the task or dataset of interest, which requires ad-hoc domain knowledge and introduces extra difficulty into algorithm designing. As such, this discrepancy presumably leads to sub-optimal fairness-aware pipelines for different applications. In this paper, we propose to fill in the fairness learning gap between these two stages by automatically learning an effective metric integrated into the fairness of both data and classifiers. Specifically, we formulate the fairness-aware classification as a distributional robustness optimization problem based on deep metric learning and propose an effective optimization algorithm to solve it. Meanwhile, we establish the asymptotically unbiased generalization bounds for the proposed algorithm using the techniques of U-statistics. The experimental results on popular benchmark datasets demonstrate that the proposed approach achieves consistent improvement concerning several fairness assessments.**

I. Introduction

In recent years we are witnessing the increasing usage of machine learning models in high-stakes decision-making, such as awarding loans, deciding probationers' risks, or detecting fraud. However, there is evidence showing that ML models can also be biased just as human decision-makers. The machine decision is either biased by the intrinsic algorithm design or twisted by the flawed collection of training data. For example, there is gender bias in Amazon's resume screening tool [1] and the credit limits of Apple Card [2]. Algorithmic fairness is gaining growing interest to address this problem. Usually, some features in the data for decision-making are also indicating the underprivileged groups in the population. Algorithmic fairness aims to learn classifiers insensitive to these features, *a.k.a.* protected attributes. For example, one trains a machine learning model to award loans based on user profiles. In this case, gender and ethnicity are protected attributes, and algorithmic fairness prevents the decision from being associated with gender or ethnicity. Some methods also require carefully handling the samples [3], [4]. Driven by different tasks, researchers proposed various definitions for algorithmic fairness, including demographic parity [5], equal odds and equal opportunities [6], disparate treatment, impact, and mistreatment [7]. Among these, individual fairness [8], based on the principle that any two individuals who are similar concerning a particular task should be classified similarly [8], is a suitable criterion for many classification problems.

A typical individually fair algorithm involves two stages. First, one needs to find a distance metric measuring the similarity between individuals. Noteworthy, the desired metric is presumably insensitive to the protected attributes [9]–[13]. Second, the algorithms learn fair classifiers [14] which are mappings from individuals to outcomes under a Lipschitz condition concerning the similarity distance metric. There are many efforts to enhance the robustness of fairness-aware models [9], [14]–[16], e.g., augmenting the labeled data [17], improving robustness to perturbations in the training distribution [18], and training against noisy protected groups [19].

In most existing methods, algorithmic fairness is characterized disjointedly by the fair metric between individuals (the data) and the fairness concerning the classifier (the learning algorithm). However, this pipeline has several important issues that remain resolved. The selection of similarity metrics is difficult but critical in individual fairness. Determining the fair metric usually involves domain knowledge associated with the specific tasks or datasets densely, which requires additional efforts in manually designing and tuning. Moreover, ignoring the relations between fair metric selection and learning algorithms often leads to sub-optimal results in practical applications. For example, the non-unified learning process requires extra effort to determine the appropriate distance metrics from a set of candidates for different application scenarios. On the other hand, manually selecting similarity metrics potentially attenuate the effectiveness of the data-driven model design. As such, the metric/classifier combinations in prior methods are seldom the optimal choices.

In this paper, we bridge this gap by learning a new metric that simultaneously optimizes the similarity assessment and the classification task. Instead of manually designed metrics, we resort to deep metrics learned from data, which have better generalization abilities and are viable to different application scenarios. More specifically, we generate perturbed instances under the fairness consideration while constraining their dissimilarity, and maximize the distance between instances with different labels. Our fair classifier utilizes the learned metric explicitly, which naturally associates their combination of the fair metric and the classifier model. On the contrary, this connection in prior works is loose, typically via a Lipschitz condition concerning the data.

Our main contributions in this paper include (1) We propose to formulate the algorithmic fairness as a distributionally

robust optimization problem. In detail, we propose to learn a deep metric to simultaneously assess the similarity of population and make a fair classification. The algorithmic fairness can be captured in our unified model, instead of in separated stages. (2) We resort to the dual problem of the original formulation and provide a solution using adversarial training. Theoretically, we derive the asymptotically unbiased generalization bound of the proposed model using M-estimator and U-statistics techniques. The analysis shows that our model generalizes almost as well as normal algorithms while admits certain fairness tolerance. (3) The experimental results on four benchmark datasets demonstrate the effectiveness of our proposed method. We also investigate some details of the proposed method.

Notations: Throughout the paper, let $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ be the feature set and $\boldsymbol{Y} = \{y_1, y_2, \ldots, y_n\}$ the label set. Let $\boldsymbol{X} \subset \mathcal{X}$ and $\boldsymbol{Y} \subset \mathcal{Y}$, and $\mathcal{X}$ and $\mathcal{Y}$ are the feature space and the label space, respectively. $r(y_i, y_j) = 1$ if $y_i \neq y_j$, otherwise $r(y_i, y_j) = 0$. Denote $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$ and $\boldsymbol{z}_i \triangleq (\boldsymbol{x}_i, y_i)$. $\mathcal{H}$ is the hypothesis class of the classifier, and $h \in \mathcal{H}$ parameterized by $\theta \in \Theta$. We use $\mathbb{E}_x$ to represent the expectation with respect to x.

II. Robust Deep Fair Learning

Prior studies typically involve two disjunctive stages, the data-wise fair metric and the model-wise fairness, which may provoke several issues. On the one hand, for a specific application, the fairness-aware data similarity usually involves domain knowledge, and existing fairness-aware learning algorithms are designed based on fixed fair similarity. As a result, some methods rely on the specific structure of the manually designed metrics and lack the flexibility for metrics with different structures. The distribution of data in practical applications is also extremely complicated, which may restrain the scope of choosing favorable metrics. On the other hand, it is presumably neglecting the data information if some learning algorithms are not dependent on the metric structures because the fairness metric is usually associated with the data distribution. Besides, it is difficult to choose the best potential learning algorithm if there are multiple candidates for the fairness data similarity.

We propose to learn a deep metric integrated into both the data-wise fair similarity and the classification fairness. In this sense, *the potential of the similarity metric is maxed out via simultaneously learning fairness for both data and algorithm*. By formulating the algorithmic fairness as a distributional robustness optimization problem integrated with a fair deep metric, the metric can not only serve in defining the protected subgroups (the instances whose protected attributes are identified as underrepresented) but also be exploited jointly in classification. Thus, we can derive an end-to-end trainable deep metric learning model with aware of fairness. We will first describe the problem of distributionally robust fairness learning, then propose our approach with theoretical analysis.

A. Problem Statement

1) Individual Fairness via Lipschitz Continuity: At a high level, algorithmic fairness can be mathematically defined by the group or by the individual. Group fairness, also referred to as statistical parity, considers the invariance of machine learning models on the protected non-overlapping subsets. Although it is compliant with statistical analysis, its prevalence is challenged due to two critical issues. First, a group-fair model is potentially blatantly unfair concerning individuals [8]. Second, many fairness constraints are intrinsically incompatible [20]. Alternatively, individually fair models are based on the intuition that similar users deserve similar treatments. It [8] views models as mapping input metric spaces to output metric spaces and defines individual fairness as Lipschitz continuity of the models. Formally, it requires:

$$d_y(h(\boldsymbol{x}_1), h(\boldsymbol{x}_2)) \leq L d_x(\boldsymbol{x}_1, \boldsymbol{x}_2), \forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}, \quad (1)$$

where d_x and d_y are metrics on the input and output spaces, h is the mapping, and $L \geq 0$ is a Lipschitz constant. Noteworthy, a preferable property of individual fairness is that the Lipschitz condition naturally implies statistical parity between subgroups of the population.

2) Fair (Wasserstein) Metric: For $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$, a distance metric $d_x(\boldsymbol{x}_1, \boldsymbol{x}_2)$, either manually designed or learned from data, is their similarity insensitive to the protected attributes. We extend the definition of $d_x(\cdot)$ to $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$ with,

$$d_z((\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2)) \triangleq d_x(\boldsymbol{x}_1, \boldsymbol{x}_2) + \infty \cdot \mathbb{I}\{y_1 \neq y_2\}, \quad (2)$$

d_z^2 is a transport cost function on $\mathcal{Z}$, which encodes the comparable samples with respect to the protected attributes. $\mathbb{I}(\cdot)$ is an indicator function.

A fair Wasserstein distance defined on the space of probability distributions on $\mathcal{Z}$ is,

$$W_c(P, Q) = \inf_{\Pi \in \mathcal{C}(P,Q)} \int_{\mathcal{Z} \times \mathcal{Z}} c(\boldsymbol{z}_1, \boldsymbol{z}_2) d\Pi(\boldsymbol{z}_1, \boldsymbol{z}_2), \quad (3)$$

where $\mathcal{C}(P, Q)$ is the set of coupling between two distributions P and Q. $c(\boldsymbol{z}_1, \boldsymbol{z}_2)$ is a fair transport cost function, *e.g.* d_z^2. The fair Wasserstein distance characterizes the similarity between sample sets, *i.e.* the distance between two probability distributions is small if they are supported on comparable areas of the sample space.

3) Distributionally Robust Fairness (DRF): For the set $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$, referred to as audit data, comparable samples can be identified using the fair Wasserstein distance. Intuitively, a predicting model is robust if the disparity of model performance on comparable samples is indistinct. To obtain a robust model, we can solve the following problem:

$$\max_{P: W_c(P, P_n) \leq \epsilon} \int_{\mathcal{Z}} \ell(\boldsymbol{z}, h) dP(\boldsymbol{z}), \quad (4)$$

where $\ell : \mathcal{Z} \times \mathcal{H} \to \mathcal{R}_+$ is a loss function, h is the prediction model, P_n is the empirical distribution of the audit data, and $\epsilon > 0$ is a small tolerance parameter. ϵ can be interpreted as a moving budget, characterizing the performance discrepancy of ML models, and forcing the evaluation on comparable areas of

samples only. Formally, the above implicit notion of fairness is summarized in the following definition.

Definition 1. *(distributionally robust fairness (DRF) [14]) A model $h : \mathcal{X} \to \mathcal{Y}$ is (ϵ, δ)-distributionally robustly fair w.r.t. the distance metric d_x iff,* $\max_{P:W_c(P,P_n)\leq\epsilon} \int_{\mathcal{Z}} \ell(\boldsymbol{z}, h) dP(\boldsymbol{z}) \leq \delta$.

Remark. *It should be emphasized that* (4) *detects an aggregate violation of individual fairness. DRF also implies individual fairness, meanwhile considers fairness-accuracy trade-off. An informal argument is that we can differ P and P_n with a single sample, and let the different samples be similar, which satisfies the Wasserstein distance constraint. DRF states that the maximum performance difference will be smaller than 2δ, which implies the existence of the Lipschitz constant for fairness. Different from the individual fairness setting, the model performance is explicitly bounded by δ.*

The aim of the fair algorithm is to learn a model with small ϵ and δ, which correspond to fairness and model accuracy, respectively. Thus, DRF can be written as, $\inf_{h\in\mathcal{H}} \max_{P:W_c(P,P_n)\leq\epsilon} \mathbb{E}_P[\ell(\boldsymbol{z}, h)]$.

B. Proposed Method

1) Robust Deep Fair Model: Existing works define the algorithmic fairness on (1), in which data-wise and algorithm-wise fairness is obtained disjointedly. However, the separated learning of the metric and the fair model could lead to sub-optimal results. Meanwhile, in practical applications, the data distribution is usually very complicated. Therefore, the deterministic a priori metric in the existing methods cannot effectively characterize the data structure.

To address these problems, we propose to learn a deep metric in the input space, not only using it as the individual fairness definition, but also using it directly in the classification. In detail, we employ an encoder to learn the representation of data, and define the following new objective based on the representation:

$$\inf_{h\in\mathcal{H}} \max_{P:W_{d_z}(P,P_n)\leq\epsilon} \mathbb{E}_{\boldsymbol{z}_1}\mathbb{E}_{\boldsymbol{z}_2,\boldsymbol{z}_3}[\ell_d(d_z(\boldsymbol{z}_1,\boldsymbol{z}_2), d_z(\boldsymbol{z}_1,\boldsymbol{z}_3), h)], \tag{5}$$

where $\ell_d : \mathcal{R}_+ \times \mathcal{R}_+ \times \mathcal{H} \to \mathcal{R}_+$ is a classification function defined also on the metric. $\boldsymbol{z}_1$ is drawn from P, $\boldsymbol{z}_2$ and $\boldsymbol{z}_3$ from P_n. (5) describes a metric loss defined on a mixed subset drawn from comparable areas, and in this paper we specify the choice of ℓ_d similar to triplet loss:

$$\ell_d(d_{12}, d_{23}, h) = [d_{12} + l - d_{23}]_+, \tag{6}$$

where $\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3$ can be viewed as generalized index anchor, positive, and negative points, respectively. h is a deep neural network embedding the input to a fair space, and for simplicity we define $d_{ij} = d_z(\boldsymbol{z}_i, \boldsymbol{z}_j) = \langle \boldsymbol{z}_1, \boldsymbol{z}_2 \rangle$ as the cosine similarity. For the implementation, we sample two data points with different labels and compute their representations $\boldsymbol{z}_2$ and $\boldsymbol{z}_3$. Then, we generate $\boldsymbol{z}_1$, a perturbed $\boldsymbol{z}_2$ sharing the same label under the fairness tolerance η. $\boldsymbol{z}_1$ denotes the biased individual which is the most different from $\boldsymbol{z}_2$ given the fairness tolerance, but belongs to the same category. We will detail the generation of $\boldsymbol{z}_1$ in the next part.

Algorithm 1: Learning A Deep Robust Fair Model

Input: data $\boldsymbol{X}, \boldsymbol{Y}$, initialized model $\boldsymbol{\theta}$, step size α, β.
Output: optimized model parameters.

repeat
  Shuffle data,
  while *epoch not end* **do**
    Get next batch from P_n
    $x^*_{t_b} \leftarrow \arg\max_{x\in\mathcal{X}} \ell_d((\boldsymbol{x}, y_{t_b}), \boldsymbol{z}_2, \boldsymbol{z}_3, \hat{\theta}_t) - \hat{\lambda}_t d_x(\boldsymbol{x}_{t_b}, \boldsymbol{x})$
    $\hat{\lambda}_{t+1} \leftarrow \max\{0, \hat{\lambda}_t - \alpha(\eta - \frac{1}{B}\sum_{b=1}^{B} d_x(\boldsymbol{x}_{t_b}, \boldsymbol{x}^*_{t_b}))\}$
    $\hat{\boldsymbol{\theta}}_{t+1} \leftarrow \hat{\boldsymbol{\theta}}_t - \frac{\beta}{B}\sum_{b=1}^{B} \partial_\theta \ell_d((\boldsymbol{x}^*_{t_b}, y_{t_b}), \boldsymbol{z}_2, \boldsymbol{z}_3, \hat{\boldsymbol{\theta}}_t)$
until *Converges.*;

Of note, the fair similarity metric d_{ij} used both as the *transport cost* and the *classification criterion* in our model, which naturally *connects the algorithmic fairness and the classifier itself*. More importantly, by a deep neural network, we can automatically learn d_{ij} from the data, which has better *generalization abilities* than the traditional fairness learning methods in practical applications.

2) Optimization Algorithm: Eq. (5) defines a distributionally robust optimization (DRO) problem. Adversarial training methods can be adopted, in light of their similarities with respect to DRO problems. Assume the hypothesis class is parametrized by $\boldsymbol{\theta} \in \Theta$, we can solve the inner problem and the outer problem in turns. The inner problem introduces an infinite-dimensional optimization problem, which can be solved by appealing to duality. We have the Lagrangian of the inner problem, and the dual problem is:

$$\begin{aligned}
&\min_{\lambda\geq 0} \max_{P} \{\lambda(\epsilon - W(P, P_n)) + \mathbb{E}_P \mathbb{E}[\ell_d(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, h)]\} \\
=&\min_{\lambda\geq 0} \left\{ \lambda\epsilon + \mathbb{E}_{P_n}\left[\max_{x\in\mathcal{X}} [\mathbb{E}_{P_n}[\ell_d((\boldsymbol{x}_4, y_1), \boldsymbol{z}_2, \boldsymbol{z}_3, \theta)] \right.\right. \\
&\left.\left. -\lambda d_x(\boldsymbol{x}_1, \boldsymbol{x}_4)] \right] \right\}.
\end{aligned} \tag{7}$$

Therefore we optimize the dual form of (5), and the problem can be written as,

$$\begin{aligned}
\inf_{h\in\mathcal{H}} \max_{P:W(P,P_n)\leq\epsilon} &\mathbb{E}_{\mathcal{C}(P,P)}[\ell_d(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, h)] = \\
&\min_{\theta\in\Theta} \min_{\lambda\geq 0} \{\lambda\epsilon + \mathbb{E}_{P_n}[\ell^c_\lambda(\boldsymbol{Z}, h)]\},
\end{aligned} \tag{8}$$

where

$$\begin{aligned}
\ell^c_\lambda(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, h) \triangleq &\max_{x\in\mathcal{X}} \mathbb{E}_{P_n}[\ell_d((\boldsymbol{x}, y_1), \boldsymbol{z}_2, \boldsymbol{z}_3, \theta)] \\
&- \lambda d_x(\boldsymbol{x}_1, \boldsymbol{x}),
\end{aligned} \tag{9}$$

The final problem (8) is amenable to stochastic optimization methods and can be solved using projected gradient descent (PGD). We describe the proposed algorithm in Algorithm 1.

For step 5 we solve a sub-problem using standard SGD. It should be mentioned the PGD we used here is principally connected with adversarial training, which is also applied to deep metric learning [21]. Informally $(\boldsymbol{x}_1, y_1)$ can be viewed as the anchor point, and solving the dual problem can be interpreted as generating a pair $(\boldsymbol{x}_3, \boldsymbol{x}_2)$ in which the anchor is similar but the prediction is unfair.

C. Theoretical Results

Our approach considers the trade-off between accuracy and fairness. More specifically, we train an encoder by solving (5), and the algorithmic fairness is bounded via slightly sacrificing the accuracy. To see this, in this section, we derive the asymptotically unbiased generalization bound of the proposed robust deep fair model, which states that our model, under certain fairness tolerance, has the asymptotically equal generalization ability as the normal classification model. We evaluate $\sup_{\boldsymbol{\theta}\in\Theta} R(W, P_n) - R(W^*, P_n)$, here $R(W, P_n) = \sup_{P:W(P,P_n)\leq\epsilon} \mathbb{E}_P\left[\ell_d(\boldsymbol{z}, h)\right] - \mathbb{E}_{P_n}\left[\ell_d(\boldsymbol{z}, h)\right]$, W^* is based on a different cost function c^*. Using the standard assumptions in DRO [14], [22], we have the main theorem.

Theorem 1. *Suppose h is a DNN with L layers and the weight parameter $\boldsymbol{W}^l (l \in [\boldsymbol{L}])$ satisfies $\|\boldsymbol{W}^l\|_F \leq B^l$. h_l is the number of the l^{th} layer. D is the bound for the feature space in Assumption 1. Let σ be a 1-Lipschitz activation function and $\|\sigma(h(\cdot))\| \leq V^L$. Let n be the sample size, u and v the number of the positive and negative labels, respectively. Under the assumptions, for any $\epsilon > 0$ with probability at least $1 - t$,*

$$\sup_{\theta\in\Theta}\left\{\left(\sup_{P:W(P,P_n)\leq\epsilon} \mathbb{E}_P\left[\ell_d(\boldsymbol{z}, h)\right] - \mathbb{E}_{P_n}\left[\ell_d(\boldsymbol{z}, h)\right]\right) - \left(\sup_{P:W(P,P^*)\leq\epsilon} \mathbb{E}_P\left[\ell_d(\boldsymbol{z}, h)\right] - \mathbb{E}_{P_n}\left[\ell_d(\boldsymbol{z}, h)\right]\right)\right\} \leq 2\delta_n, \tag{10}$$

here, $\delta_n \leq \frac{L\delta_c}{\sqrt{\epsilon}} + \sqrt{\frac{1}{2}n\phi^2 \ln\frac{1}{t}} + 6\sqrt{\frac{1}{\lfloor\frac{n}{3}\rfloor}} + 64h_L B^L V^L(\sqrt{2L\log 2} + 1))(\Pi_{l=1}^L B^l)\sqrt{\frac{d}{\lfloor\frac{n}{3}\rfloor}}$ *and* $\phi = \frac{2(nv+nu+vu)}{n(n-1)(n-2)}(B^L V^L D^2 + 1)$.

Remark. *The final result indicates that the generalization bound of the proposed method to a potential transport cost function is characterized by $2\delta_n$.*

III. Experimental Results

The experiments aim to demonstrate that our method achieves comparable or superior fairness compared to related fair algorithms. To support our claim, the experiments are conducted on four standard datasets concerning several representative fairness metrics.

Experimental Setting: We focus on four standard tasks: income prediction, recidivism prediction, credit risk prediction, and deposit prediction [23]. We compute three different fair metrics for the comparison with baselines. Let $\mathcal{C}$ be a set of classes, A be a binary protected attribute, and $Y, \hat{Y} \in \mathcal{C}$ be the true and predicted class labels, respectively. For simplicity let $a \in \{0, 1\}$ and $c \in \mathcal{C}$. We include Log-probabilistic equalized opportunities *LogUNF* [24], *Consistency*, and True Positive Rate *TPR Gap* as the fair metrics and define them in the following. The accuracy is also included to show the trade-off for algorithmic fairness.

LogUNF: Let a be the binary protected attribute, $\boldsymbol{x}$ the features, y the label, P the distribution of data. LogUNF is defined as $|\mathbb{E}_P[1 + \log h(\boldsymbol{x})|a = 1, y = 1] - \mathbb{E}_P[1 + \log h(\boldsymbol{x})|a = 1, y = 1]|$.

Consistency: Given a binary protected attribute, we make two copies of every data points by flipping the value of the protected attribute. We define $\hat{Y}_a$ as the predicted labels, with the protected attribute set to a, and the consistency can be defined as $Con = P(\hat{Y}_1 = \hat{Y}_0)$.

TPR Gap: We define $TPR_{a,c} = P(\hat{Y} = c|A = a, Y = c)$, $Gap_{A,c} = TPR_{0,c} - TPR_{1,c}$, and the rooted mean square and the max of the gap are considered, defined as $Gap_A^{RMS} = \sqrt{\frac{1}{|C|}\sum_{c\in C} Gap_{A,c}^2}$ and $Gap_A^{max} = \arg\max_{c\in C} |Gap_{A,c}|$.

We compare the proposed method with several related methods, including baseline, DAML [21], Project [14], CoCL [25], Adversarial Debiasing [26], SenSR [14], and EXPLORE [9]. We use the results from the original paper when available, and the rest results of baselines are obtained using the implementation provided by the authors. For the proposed method, we use a single layer neural network with 100 hidden units and 64-dim outputs to compute the data representations, compute the inner-product between instances as their distance in the training, and use k nearest neighbor on the representations for the classification with $k = 5$ in the inference. α and β are set to 0.001, and ϵ is set to 0.01. $l = 1$ for the margin in the triplet loss. The model is trained 500 epochs using Adam optimizer with a learning rate of 0.0001. All experimental results are obtained on a single Nvidia P40 GPU.

Numerical Results: We present the numerical results for the involved tasks in this part and discuss the observations concerning the results. For all experiments, we split the dataset, using 80% as the train set and 20% as the test set. The results are obtained by averaging the fairness metrics on the test sets based on ten random splits. The descriptions and the setting of the tasks are included in the following discussions.

Income Prediction The Adult dataset [27] is from the Census Bureau and the task is to predict whether a given adult makes more than $\$50,000$ a year based on attributes such as education, hours of work per week, etc., for approximately $45,000$ individuals. In this experiment, *gender* (male or female) and *race* (Caucasian or non-Caucasian) are used as binary protected attributes. The computational results are presented in Table I.

Recidivism Prediction Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a commercial tool to assess a criminal defendant's likelihood to re-offend, containing 5278 instances. The task is to predict the recidivism risk based on the features for defendants, including the criminal history, jail and prison time, demographics, etc. In this experiment, *gender* (male or female) and *race* (Caucasian

or non-Caucasian) are used as binary protected attributes. The computational results are presented in Table II.

Credit Risk Prediction The German dataset contains the credit data of 1000 instances, and the task is to predict the credit risks. In this experiment, *age* as the protected attribute. The results are summarized in Table III.

Deposit Prediction The Bank dataset is collected by a Portuguese banking institution, and the task is to predict if the client will subscribe to a term deposit. This dataset contains 41188 examples, and we use 30488 examples in this experiment by discarding those with missing records. In this experiment, *gender* as the protected attribute. The results are summarized in Table IV.

Discussions: In this experiment, we have several observations. For baselines, almost all fairness criteria are violated in the Adult data; in COMPAS, TPR Gap is tolerable. A naive deep metric method, DAML, exhibits similar property. Noteworthy, DAML usually performs better on TPR Gap, compared to the baseline. On the other hand, fair algorithms generally show effectiveness within their scope of fairness. For example, SenSR and the related variants show significant improvements on the Adult dataset, particularly on the consistency concerning the protected attributes. We also witness the unstable performance of these methods. For example, Project works well concerning consistency, but on different datasets, the TPR Gap perturbs severely, which puts the practical usage in doubt when it is extended to novel tasks.

The proposed method demonstrates its effectiveness. First, the proposed method can make predictions with better consistency compared to related methods. Second, the proposed method can significantly improve the gap of true positive rate concerning the protected attributes. Third, the proposed method can attain decent fairness in all senses compared to the related methods. The most prominent advantage is that the proposed method achieves the best LogUNF and TPR gap among all methods, which is illustrated in Fig. 1. The proposed method consistently outperforms related methods, particularly for the TPR gap. A potential explanation is that the triplet loss can learn a more powerful metric. Intuitively, the metrics used by previous methods mainly consider the local similarity, which only involves a small subset of samples. The metric learned by the proposed method not only considers the local similarity but also considers the global similarity for the classification, which strengthens the fairness.

Ablation Study Compared to previous related methods, our approach takes longer computation time. E.g. SenSR and its variants also involve the distributionally robust optimization. Our method includes an additional step to construct the pairs of different labels. A comparison is included in Table V.

Our method is trained under the supervised learning framework. The final classification is obtained via k-nearest neighbors. In this part we examine the effect of different k. The results are summarized in Table. VI. The results show that moderate k is sufficient for good performance, and k is not very sensitive. When k grows large (*i.e.* ≥ 8), the results may degenerate, particularly for accuracy.

IV. Conclusion

In this paper, we formulate the fair classification as a distributionally robust optimization problem, guided by a learned metric. To bridge the gap in previous two-stage pipelines, we propose a new robust deep fair model to learn a metric function jointly in assessing input similarity and fair classification. We derive the generalization bound using the U-statistics techniques. The experimental results demonstrate the effectiveness of our proposed method on benchmark datasets.

References

[1] J. Dastin, "Amazon scraps secret ai recruiting tool that showed bias against women," *Reuters*, 2018.

[2] N. Vigdor, "Apple card investigated after gender discrimination complaints," *The New York Times*, 2019.

[3] R. Bao and Others, "Efficient approximate solution path algorithm for order weight l_1-norm with accuracy guarantee," in *2019 IEEE International Conference on Data Mining (ICDM)*, 2019.

[4] ——, "Fast oscar and owl regression via safe screening rules," in *International Conference on Machine Learning*, 2020.

[5] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 2009, pp. 13–18.

[6] M. Hardt and Others, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016.

[7] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1171–1180.

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

[9] D. Mukherjee, M. Yurochkin, M. Banerjee, and Y. Sun, "Two simple ways to learn individual fairness metrics from data," *arXiv preprint arXiv:2006.11439*, 2020.

[10] A. Bower, L. Niss, Y. Sun, and A. Vargo, "Debiasing representations by removing unwanted variation due to protected attributes," *arXiv preprint arXiv:1807.00461*, 2018.

[11] M. Donini and Others, "Empirical risk minimization under fairness constraints," in *Advances in Neural Information Processing Systems*, 2018, pp. 2791–2801.

[12] S. Gillen and Others, "Online learning with an unknown fairness metric," in *Advances in neural information processing systems*, 2018.

[13] C. Jung and Others, "Eliciting and enforcing subjective individual fairness," *arXiv preprint arXiv:1905.10660*, 2019.

[14] M. Yurochkin, A. Bower, and Y. Sun, "Training individually fair ml models with sensitive subspace robustness," in *International Conference on Learning Representations*, 2019.

[15] F. Yang, M. Cisse, and O. O. Koyejo, "Fairness with overlapping groups; a probabilistic perspective," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[16] H. Xu, X. Liu, Y. Li, and J. Tang, "To be robust or to be fair: Towards fairness in adversarial training," *arXiv preprint arXiv:2010.06121*, 2020.

[17] D. Ji, P. Smyth, and M. Steyvers, "Can i trust my fairness metric? assessing fairness with unlabeled data and bayesian inference," *arXiv preprint arXiv:2010.09851*, 2020.

[18] D. Mandal, S. Deng, S. Jana, J. M. Wing, and D. Hsu, "Ensuring fairness beyond the training data," *arXiv preprint arXiv:2007.06029*, 2020.

[19] S. Wang, W. Guo, H. Narasimhan, A. Cotter, M. Gupta, and M. I. Jordan, "Robust optimization for fairness with noisy protected groups," *arXiv preprint arXiv:2002.09343*, 2020.

[20] "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

[21] Y. Duan, W. Zheng, X. Lin, J. Lu, and J. Zhou, "Deep adversarial metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2780–2789.

[22] J. Lee and Others, "Minimax statistical learning with wasserstein distances," in *Advances in Neural Information Processing Systems*, 2018.

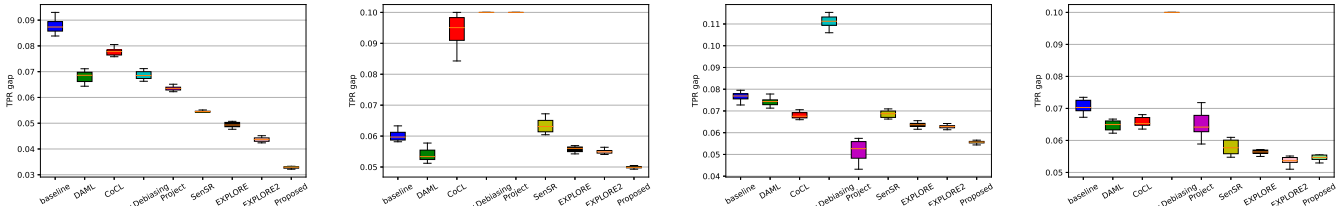
[23] D. Dua and C. Graff, "Uci machine learning repository," 2017.

TABLE I: Computational results on Adult datasets. The best results are shown in bold font. S-Con and GR-Con refer to the spouse-consistency and the gender/race consistency, respectively. Gap_R and Gap_G refer to the race and the gender TPR gap. Accuracy is in %

	Accuracy	LogUNF	S-Con	GR-Con	Gap_G^{RMS}	Gap_R^{RMS}	Gap_G^{max}	Gap_R^{max}
Baseline	82.9	0.95	0.848	0.865	0.179	0.089	0.216	0.105
DAML	81.7	0.76	0.810	0.833	0.083	0.071	0.116	0.080
CoCL	79.0	–	–	–	0.163	0.080	0.201	0.109
Adv. debiasing	81.5	0.66	0.807	0.841	0.082	0.070	0.110	0.078
Project	82.7	0.70	0.868	**1.00**	0.145	0.064	0.192	0.086
senSR+FACT	78.9	0.58	0.934	0.984	0.068	0.055	0.087	0.067
senSR+Explore (no gender)	78.9	0.57	0.933	0.993	0.066	0.050	0.084	0.063
senSR+Explore (gender)	79.4	0.56	0.966	0.987	0.065	0.044	0.084	0.059
Our Method	80.4	**0.51**	**0.970**	0.995	**0.047**	**0.033**	**0.066**	**0.043**

TABLE II: Computational results on COMPAS dataset.

	Accuracy	LogUNF	G-Con	R-Con	Gap_G^{RMS}	Gap_R^{RMS}	Gap_G^{max}	Gap_R^{max}
Baseline	53.8	0.26	0.832	0.840	0.060	0.042	0.066	0.054
DAML	62.5	0.23	0.805	0.812	0.054	**0.041**	0.058	**0.051**
Adv. debiasing	65.2	0.22	0.785	0.793	0.283	0.227	0.337	0.254
Project	65.7	0.22	0.876	0.886	0.136	0.181	0.168	0.209
senSR+FACT	57.6	0.15	0.870	0.892	0.064	0.053	0.081	0.066
senSR+Explore (no gender)	56.9	0.12	0.884	0.893	0.056	0.045	0.072	0.057
senSR+Explore (gender)	56.7	0.11	0.878	0.902	0.055	0.044	0.071	0.055
Our Method	61.9	**0.09**	**0.891**	**0.909**	**0.050**	**0.041**	**0.054**	**0.051**



(a) Adult (b) COMPAS (c) German (d) Bank

Fig. 1: Variance of TPR gap.

TABLE III: Computational results on German dataset. Explore* refers to EXPLORE using the protected attribute.

	Accuracy	LogUNF	Gap^{RMS}	Gap^{max}
Baseline	72	0.32	0.077	0.109
DAML	73	0.31	0.075	0.105
Adv. debiasing	70	0.27	0.112	0.149
Project	70	0.24	0.057	0.081
SenSR	68	0.15	0.069	0.098
Explore	69	0.14	0.064	0.092
Explore*	68	0.14	0.063	0.092
Our Method	71	**0.12**	**0.055**	**0.080**

TABLE IV: Computational results on Bank dataset.

	Accuracy	LogUNF	Gap^{RMS}	Gap^{max}
Baseline	75	0.23	0.051	0.071
DAML	74	0.24	0.047	0.065
Adv. debiasing	71	0.22	0.089	0.122
Project	72	0.19	0.051	0.068
SenSR	74	0.18	0.043	0.058
Explore	74	0.16	0.043	0.057
Explore*	74	0.16	**0.042**	0.054
Our Method	76	**0.14**	**0.042**	**0.053**

TABLE V: Per-epoch computational time (in second) of different methods for two datasets.

	Baseline	SenSR	Explore	Our Method
Adult	6.9	25.1	23.4	29.8
German	0.8	3.4	3.5	6.1

TABLE VI: Per-epoch computational time of different methods for Bank dataset.

k	Accuracy	LogUNF	Gap^{RMS}	Gap^{max}
3	76	0.16	0.045	0.060
4	75	0.14	0.044	0.057
5	76	0.14	0.042	0.053
6	76	0.15	0.041	0.053
8	75	0.14	0.042	0.055
10	73	0.13	0.044	0.056

[24] B. Taskesen and Others, "A distributionally robust approach to fair classification," *arXiv preprint arXiv:2007.09530*, 2020.

[25] M. De-Arteaga and Others, "Bias in bios: A case study of semantic representation bias in a high-stakes setting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

[26] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[27] R. Kohavi, "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid." in *Kdd*, vol. 96, 1996, pp. 202–207.

ACKNOWLEDGMENT

This work was partially supported by NSF IIS 1845666, 1852606, 1838627, 1837956, 1956002, 2040588.