IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 5, 2020, accepted June 16, 2020, date of publication June 30, 2020, date of current version July 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3005987

Joint Transfer of Model Knowledge and Fairness Over Domains Using Wasserstein Distance

TAEHO YOON[1], JAEWOOK LEE[2], (Member, IEEE), AND WOOJIN LEE[2]
[1]Department of Mathematical Sciences, Seoul National University, Seoul 08826, South Korea
[2]Department of Industrial Engineering, Seoul National University, Seoul 08826, South Korea

Corresponding author: Woojin Lee (wj926@snu.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Korean Government Ministry of Science and ICT (MSIT) under Grant NRF-2019R1A2C2002358 and Grant 2017R1A5A1015626.

ABSTRACT Owing to the increasing use of machine learning in our daily lives, the problem of fairness has recently become an important topic in machine learning societies. Recent studies regarding fairness in machine learning have been conducted to attempt to ensure statistical independence between individual model predictions and designated sensitive attributes. However, in reality, cases exist in which the sensitive variables of data used for learning models differ from the data upon which the model is applied. In this paper, we investigate a methodology for developing a fair classification model for data with limited or no labels, by transferring knowledge from another data domain where information is fully available. This is done by controlling the Wasserstein distances between relevant distributions. Subsequently, we obtain a fair model that could be successfully applied to two datasets with different sensitive attributes. We present theoretical results validating that our approach provably transfers both classification performance and fairness over domains. Experimental results show that our method does indeed promote fairness for the target domain, while retaining reasonable classification accuracy, and that it often outperforms comparative models in terms of joint fairness.

INDEX TERMS Fair machine learning, fair classification, demographic parity, equal opportunity, domain adaptation, transfer learning.

I. INTRODUCTION

Machine learning is now widely used in a variety of decision-making scenarios such as health care, criminal risk assessment, and financial lending. As machine learning is permeating our everyday lives, its fairness is becoming a real issue, and researchers are investigating the problem. Machine learning models are trained to predict outcomes for new samples using a given set of labeled examples. Because the process relies significantly on the selected dataset, and the training data might possess intrinsic bias toward historically discriminated groups, there has been emerging concern that models can inherit and reproduce such unfair treatments.

Traditional machine learning models for classification were designed to maximize the accuracy of their predictions; however, accurate predictions may still be unfair. This led to the growth of the literature on fairness in machine learning. Usually, these works consider one of the following two types: individual or group fairness. Individual fairness [1] claims that similar users should be treated similarly, while group fairness, which we consider in this study, attempts to obtain decisions that are fair in the sense that the outcomes should not allow inferences on the specific group information [2]. Such group information includes variables such as gender, age, and race, which are known as ''sensitive attributes.''

Demographic parity (DP), or *statistical parity*, is one of the most widely used criterion for fairness. It requires statistical independence between model predictions and sensitive attributes [3]. That is, DP is achieved when the ratio of each output is equal for the set of inputs belonging to each sensitive group. For example, in a financial lending problem in which the sensitive attribute of choice is gender, lending approval rates for male and female groups should be the same.

However, when the base rates between sensitive groups differ significantly, DP might not completely encapsulate the concept of fairness. *Equal opportunities (EOp)* has been suggested as an alternative fairness notion for binary classification setting under such circumstances [4]. It only enforces

The associate editor coordinating the review of this manuscript and approving it for publication was Long Cheng.

non-discrimination over the "advantaged" outcome, i.e., it requires equal true positive rates for each sensitive group. Hence, achieving fairness in terms of EOp may result in more accurate, reasonable predictions.

DP and EOp both enjoys popularity as fairness metrics, but they have several technical differences. In the training phase, minimizing EOp violation requires access to label information, whereas DP does not. Some researchers insist that the two criteria are incompatible, except in special cases wherein the sensitive attribute is independent of the target variable [5]. Therefore, it is important to set an appropriate fairness goal compatible with the setting. In this study, we considered both DP and EOp as fairness criteria.

Several distinct streams of works have been done to develop machine learning models with fair results. The first approach directly enforced the designated fairness criterion as constraints during model training, for example, by using kernel methods or by reducing the Wasserstein distance [6]–[9]. Another major approach was to search for representations of data in which information pertaining to sensitive attributes is removed [10]–[12]. Some recent works along this trend have applied adversarial training to generate similar data representations for different sensitive groups, so that any classifier acting on the representations will be agnostic [12]. The method that we propose is related to both approaches, as it trains a fair classifier by learning appropriate latent representations.

Previous studies regarding fair classification were mainly focused on removing DP or EOp for a single sensitive attribute. However, in reality, models could be applied to multiple tasks, each with a distinct sensitive attribute of interest. Additionally, the sensitive attribute in labeled training data may differ from that in the target data with limited or no labels. For example, if one is provided with a substantial amount of labeled data with sensitive attribute *race*, but the trained model should also be applied to a similar dataset with *gender* as sensitive attribute, then traditional fair classification algorithms would fail to achieve fairness in terms of *gender*.

In this study, we investigate a methodology to develop fair classification models that can be applied to datasets with different sensitive attributes, while using the label information from only one dataset. This problem is equivalent to constructing a model that achieves fairness over multiple sensitive attributes from a dataset with limited label access. To that end, we exploit and present ideas stemming from the intersection of domain adaptation and fair machine learning literature.

Domain adaptation is an aspect of transfer learning that attempts to train a machine learning model from labeled source data, so that it performs well on similar, but different, target data. It is assumed that the source and target domains are associated to the same classification task, but have different underlying distributions. Our research question could be regarded as a domain adaptation problem in which the source and target domains have distinct sensitive attributes. We have labeled source data and unlabeled target data, and the goal is to construct a model that ensures fairness in terms of both sensitive attributes.

Figure 1 summarizes the experimental results for our method, in the form of confusion matrices. The acceptance rate for each sensitive group is shown at the bottom of the matrices, and the difference between acceptance rates measure the degree of violation of DP. Figure 1a shows the results obtained by a previous fair classification method [7]. The DP gap with respect to the sensitive attribute in the source domain (*race*) is successfully minimized ($\Delta\text{DP}race = |0.121 - 0.130| = 0.009$); however when applied to the sensitive attribute in the target domain (*gender*), it fails to reduce the DP gap ($\Delta\text{DP}gender = |0.040 - 0.166| = 0.126$). In contrast, our proposed method effectively reduces the DP gap in both the target ($\Delta\text{DP}gender = |0.117 - 0.126| = 0.009$) and the source domains ($\Delta\text{DP}race = |0.100 - 0.112| = 0.012$), as shown in Figure 1b.

We rewrite the fairness metric in terms of Wasserstein distance, and combine this with domain adaptation techniques based on Wasserstein distance to produce training objectives that can be computed efficiently. This could be done by replacing Wasserstein distance between multi-dimensional distributions with sliced Wasserstein discrepancy, and taking advantage of a simple closed-form expression for Wasserstein distance for certain types of one-dimensional distributions. It turns out that usage of Wasserstein-based objective makes our method capable of addressing the disparity of the classifier for virtually all threshold values τ, by matching the cumulative distribution of score values in both the source and target domains.

In Section 2, we provide a brief review of previous studies related to domain adaptation and fair classification. In Section 3, we present the theory behind the formulation of our method and its actual implementation. In Section 4, we verify the effectiveness of our method through experiments on three real datasets. Finally, in Section 5, we discuss the contributions of our study and future works.

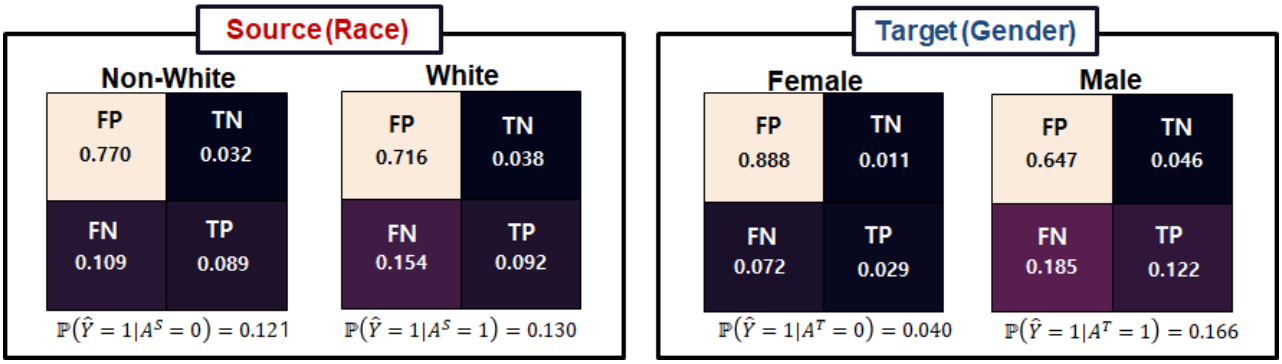
II. RELATED WORK

In this section, we review previous studies related to fair classification and domain adaptation. Additionally, we briefly mention methods similar to ours for comparison to the proposed method in later sections.

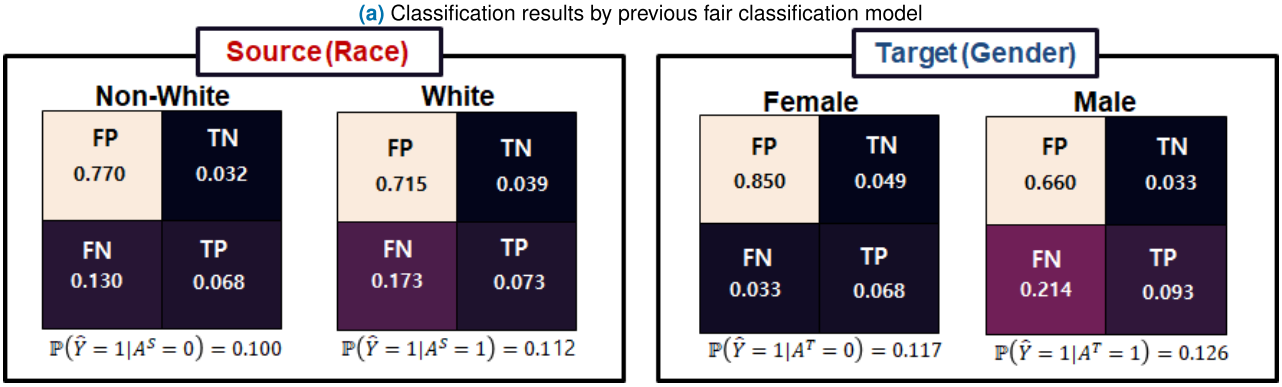
A. FAIRNESS

In a general fair classification problem, we consider training observations (X, A, Y), where $X \in \mathbb{R}^d$ is an input or feature vector, $Y \in \{0, 1\}$ is the label, and $A \in \{0, 1\}$ is the sensitive attribute. The main purpose of fair classification is to learn $\hat{Y} \in \{0, 1\}$ to accurately predict the true label Y, while maintaining fairness with respect to sensitive attribute A.

The fairness criterion of DP enforces statistical independence between the predicted outcome $\hat{Y}$ and the sensitive attribute A, i.e., $\hat{Y} \perp A$. Therefore, the model satisfies DP,



(a) Classification results by previous fair classification model



(b) Classification results by our method

FIGURE 1. **Confusion matrix for a previous method and our proposed method, applied to domains with different sensitive attributes. Training was done using labeled source data and unlabeled target data. Previous fairness method minimized the DP gap for the source domain sensitive attribute (ΔDP*race*:0.009), but failed to transfer fairness to the target domain (ΔDP*gender*:0.126). In contrast, our method successfully reduced the DP gap for both sensitive attributes (ΔDP*race*:0.012; ΔDP*gender*:0.009).**

if the following holds: $\Pr_X(\hat{Y} = 1|A = 0) = \Pr_X(\hat{Y} = 1|A = 1)$. This amounts to saying that the sensitive-group-conditional acceptance rates are equal. Despite its wide use, DP has shown limitations in many supervised tasks [4], e.g., it does not necessarily produce "reasonable" predictions. Even if we randomly select individuals to be accepted in one group, regardless of their true labels (i.e., whether they deserve acceptance), DP is achieved whenever the acceptance percentages match. Hence, the alternative criterion EOp has been suggested [4]. A binary predictor $\hat{Y}$ satisfies EOp if $\Pr_X(\hat{Y} = 1|A = 0, Y = 1) = \Pr_X(\hat{Y} = 1|A = 1, Y = 1)$. It focuses on matching the acceptance rates across groups on the "advantaged" outcome ($Y = 1$).

Imposing the aforementioned fairness constraints is known to conflict with learning a well-calibrated classifier [11]. For instance, when the true label depends on the sensitive attribute, DP would be incompatible with the ideal prediction. Thus, it is important to build a model satisfying the fairness requirements, while minimizing utility losses.

There has been vast literature on algorithmic approaches to fair classification. A number of works have been devoted to directly optimizing the classifier by imposing fairness constraints [7]–[9], or seeking for fair representations by solving appropriate minimax problems [10]–[12]. Some researchers have attempted to prepossess (or repair) the input in order to remove the disparate impact [2], [3].

Among various works on fair classification, we have been inspired primarily by methods related to Wasserstein distance. Fairness with respect to some sensitive attribute is closely related to similarity between conditional distributions for each sensitive group, and Wasserstein distance has been suggested as an associated discrepancy measure in several works, for it reflects the metric on the sample space [2], [7].

In some works, the Wasserstein metric has been used to repair the input data to achieve fairness [2], [3]. Gordaliza *et al.* [2] theoretically demonstrated that the Wasserstein metric is a natural choice for measuring the distance between conditional distributions in fairness problems. Based on finding Wasserstein barycenters of the distributions, geometric repair [3] and random repair [2] methods have been proposed. However, these repairing methodologies are based on solving a costly linear program, and they do not provide any guidelines for optimizing a classifier for a specific task.

Jiang *et al.* [7] suggested model training via logistic regression with 1-Wasserstein distance penalty for group-conditional score distributions, based on gradient descent. Another study used 2-Wasserstein distance in training neural-network based models, in which optimization was done using gradient approximation [8].

In this paper, we attempt to learn a model that is fair with respect to sensitive attributes from both source and target domains, while retaining its classification performance

by taking advantage of the properties of Wasserstein distance.

B. DOMAIN ADAPTATION

The basic assumption of domain adaptation is that the training (source) and test (target) data are from similar, but different, distributions. There are two scenarios in domain adaptation that are actively studied: unsupervised and semi-supervised settings, which are categorized based on whether the target data is completely unlabeled or partially labeled. Because unsupervised domain adaptation involves more challenging and general conditions, it has been more widely investigated.

There have been two main approaches developed to address domain adaptation problems. The first solution is instance-based learning [13], [14], which focuses on training a classifier that takes into account the difference between domains. The second solution utilizes the idea of representation learning [15]–[20], and it seeks transferable representations for the source and the target domains. The latter approach, which is compatible with deep neural networks, has been studied extensively in recent years.

Theoretical analysis shows that the empirical risk of the target domain can be bounded using the risk of the source domain and the $\mathcal{H}$-divergence between two domains [21]. Therefore, by minimizing $\mathcal{H}$-divergence between two domains, one can reduce the gap between empirical risk values. This can be achieved by rendering indistinguishable the representations of the source and the target domain data in the latent space. Based on the theory, several recent studies have focused on obtaining domain-invariant representations using maximum mean discrepancy (MMD) [15] or adversarial training [16]–[18]. In [22], using Wasserstein distance in adversarial training was suggested to minimize the dissimilarity between the source and target domain distributions.

So far, few studies have considered transfer of a model's knowledge on fairness to other domains. In [23], a fair transfer learning problem was addressed via instance-based domain adaptation technique, under the assumption that the sensitive attribute for only one of either the source or the target domain was available. Madras *et al.* [11] partially considered the problem by testing whether the fairness of their method (LAFTR) was preserved over different tasks.

Schumann *et al.* [24] mainly investigated the transferability of a fair model for different sensitive attributes. This is the study that is closest to ours in terms of the problem definition. They handled the problem in which the sensitive attributes differ over the source and the target domains, but the prediction tasks were the same. By extending the general theory of domain adaptation from [21], they developed a representation-learning-based method, using MMD regularizers to minimize $\mathcal{H}$-divergence between the two domains.

Unlike in previous studies [21], [24], we advocate the employment of Wasserstein distance when building a transferable model, instead of $\mathcal{H}$-divergence, which is difficult to deal with directly and requires approximation using MMD [15], [24], [25] or Jensen–Shannon divergence [11], [17]. In the next section, we present a new generalization bound for the fair domain adaptation problem using the Wasserstein metric. Subsequently, we propose our algorithm to reduce both the upper bound on the risk and the disparity in the target domain.

III. METHOD

A. THEORETICAL BACKGROUND

1) NOTATION

We consider classification tasks in which $\mathcal{X} \subset \mathbb{R}^d$ is an input space and $h : \mathcal{X} \to \{0, 1\}$ is a binary classifier that assigns either *success* (which corresponds to 1) or *failure* (which corresponds to 0). Binary class prediction h is determined by score function $\eta : \mathcal{X} \to [0, 1]$, which estimates the probability of success for each sample x, i.e., the probability that the true label associated to the sample is 1. The dataset consists of samples drawn from the joint distribution of (X, A, Y), where $X \in \mathcal{X}$ denotes the data, $A \in \{0, 1\}$ the sensitive attribute, and $Y \in \{0, 1\}$ the label.

Given a classifier h, the DP gap with respect to sensitive attribute A is defined as

$$\Delta\mathrm{DP}(h) = \left| \Pr_X(\hat{Y} = 1|A = 1) - \Pr_X(\hat{Y} = 1|A = 0) \right|. \quad (1)$$

Similarly, the EOp gap is given by

$$\Delta\mathrm{EOp}(h) = \left| \Pr_X(\hat{Y} = 1|Y = 1, A = 1) - \Pr_X(\hat{Y} = 1|Y = 1, A = 0) \right| \quad (2)$$

In our setting, we consider a family of thresholded classifiers $\{h_\tau\}_{\tau\in(0,1)}$ with the classification rule

$$h_\tau(x) = 1_{\{\eta(x)>\tau\}}(x),$$

i.e., h_τ predicts success if and only if the score value η exceeds τ. We often denote $h_\tau(X)$ as $\hat{Y}_\tau$ under this setting. Based upon these concepts, we introduce the notion of Strong Pairwise Demographic Disparity (SPDD), which has been originally proposed by [7], and Strong Pairwise Disparity of Opportunity (SPDOp), which is a corresponding concept for the criterion EOp. Roughly speaking, they measure the averaged gap of (conditional) success probabilities across groups, over the family of classifiers under consideration.

Definition 1: Let $\eta : \mathcal{X} \to [0, 1]$ be a score function, and let $\hat{Y}_\tau = h_\tau(X)$. Then we define SPDD and SPDOp associated to η respectively as

$$\mathrm{SPDD}(\eta) = \mathbb{E}_{\tau\sim U([0,1])}\Delta\mathrm{DP}(h_\tau)$$
$$\mathrm{SPDOp}(\eta) = \mathbb{E}_{\tau\sim U([0,1])}\Delta\mathrm{EOp}(h_\tau).$$

In our scenario, we consider the source and target data distributions on $\mathcal{X}$, each having a joint relationship to a distinct sensitive attribute. The sample distribution underlying the source data is denoted as $\mathcal{D}^S$, and that of the target data is $\mathcal{D}^T$.

Likewise, the sensitive attribute for each data is expressed as A^S and A^T.

2) OPTIMAL TRANSPORT AND WASSERSTEIN DISTANCE

For the sake of completeness, we briefly introduce some preliminary facts about Wasserstein distance.

Wasserstein distance is a measure of discrepancy between two probability distributions based on a binary cost function on the sample space. Let $\mathcal{Z}$ be the latent space and fix a binary function $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$. Given two probability measures ν_0 and ν_1 on $\mathcal{Z}$ satisfying $\int c(x, y)\, d\nu_i < \infty$ for all $y \in \mathcal{Z}$ and $i = 0, 1$, the (Monge) optimal transport problem attempts to find a transport map $T : \mathcal{Z} \to \mathcal{Z}$ that minimizes the total transport cost

$$\int_{\mathcal{Z}} c(z, T(z))\, d\nu_0(z)$$

under the condition $T_{\#}\nu_0 = \nu_1$, meaning that T push-forwards ν_0 to ν_1. The optimal transport (OT) map T^* is the one that minimizes the above quantity.

Kantorovitch [26] gave a generalized formulation of the optimal transport problem in terms of joint probability distributions:

$$\text{minimize} \int_{\mathcal{Z}\times\mathcal{Z}} c(z_0, z_1)\, d\gamma(z_0, z_1)$$

over $\gamma \in \prod(\nu_0, \nu_1) = \{\gamma \mid \pi_{i\#}\gamma = \nu_i,\ i = 0, 1\}$, where π_i's denote marginal projections induced by the canonical projections $\mathcal{Z} \times \mathcal{Z} \to \mathcal{Z}$. An optimal solution γ^* for the problem is called an optimal coupling.

Finally, suppose that $\mathcal{Z}$ has a metric space structure d. In this case, for $p \geq 1$, we define the p-Wasserstein distance as

$$W_p(\nu_0, \nu_1) = \inf_{\gamma \in \prod(\nu_0, \nu_1)} \left(\int_{\mathcal{Z}\times\mathcal{Z}} d(z_0, z_1)^p d\gamma(z_0, z_1) \right)^{\frac{1}{p}}. \quad (3)$$

Note that, except the $(1/p)$-th power, this is a special case of the previous definition with the choice $c(z_1, z_2) = d(z_1, z_2)^p$ of cost function. An advantage of considering p-Wasserstein distances is that they are indeed "distances" between probability measures, i.e., positivity, reflexivity, and triangular inequality are satisfied.

For the 1-Wasserstein distance W_1, there is a useful equivalent characterization due to Villani [27]:

$$W_1(\nu_0, \nu_1) = \sup \left\{ \int_{\mathcal{Z}} f\, d\nu_0 - \int_{\mathcal{Z}} f\, d\nu_1 \ :\ \|f\|_L \leq 1 \right\},$$

wherein the condition $\|f\|_L \leq 1$ requires f to be 1-Lipschitz as a function from $\mathcal{Z}$ to $\mathbb{R}$ with respect to the metric d, i.e., $|f(z_0) - f(z_1)| \leq d(z_0, z_1)$ for all $z_0, z_1 \in \mathcal{Z}$.

Wasserstein distance has recently gained popularity as an ingredient for loss functions in the field of artificial intelligence, due to its advantage over other discrepancy measures between probability distributions, such as total variation distance, Kullback-Leibler divergence, and Jensen-Shannon divergence. References [22], [28]–[30]. Since Wasserstein distance takes into account the properties of the underlying geometry, unlike the other dissimilarity measures mentioned, it assigns finite distance value even when two distributions do not share support [28]. Moreover, convergence with respect to the topology induced by Wasserstein distance coincides with convergence in distribution [31].

3) FAIRNESS IN TERMS OF WASSERSTEIN DISTANCE

In this subsection, we express the fairness criteria SPDD and SPDOp in terms of Wasserstein distance as in [7], which have originally proposed the idea. Then, we illustrate our framework based on the reformulation and provide a theoretical bound on the disparity between groups.

We first state a well-known fact regarding Wasserstein distance between distributions on the unit interval $\Omega = [0, 1]$. Suppose that R_0 and R_1 be random variables taking values in Ω, and for $i = 0, 1$, let F_i be the cumulative distribution function of R_i. If μ_i be the distribution on Ω induced by the variable R_i, then we have

$$W_1(\mu_0, \mu_1) = \int_{\Omega} |F_0(\tau) - F_1(\tau)|\, d\tau, \quad (4)$$

provided that μ_i's have density functions. (For proof, we refer the reader to [7].)

Importantly, we observe that (4) is directly related to the Strong Pairwise Disparity concepts of fairness under the right context. Consider the sensitive group-conditional source distributions $\mathcal{D}_a^S = \mathcal{L}(\mathcal{D}^S \mid A^S = a)$, where $a = 0, 1$. Suppose then that we have a trained score function $\eta : \mathcal{X} \to [0, 1]$, and denote the push-forwarded distributions by

$$\mu_a^S := \eta_{\#}\mathcal{D}_a^S, \qquad a = 0, 1.$$

Then, by applying (4), we obtain

$$\begin{aligned}
&W_1(\mu_0^S, \mu_1^S) \\
&= \int_0^1 \left| \Pr_{X\sim\mathcal{D}_0^S}(\eta(X) \leq \tau) - \Pr_{X\sim\mathcal{D}_1^S}(\eta(X) \leq \tau) \right| d\tau \\
&= \int_0^1 \left| \Pr_{X\sim\mathcal{D}_0^S}(\eta(X) > \tau) - \Pr_{X\sim\mathcal{D}_1^S}(\eta(X) > \tau) \right| d\tau \\
&= \int_0^1 \left| \Pr_X(\hat{Y}_\tau = 1 \mid A^S = 0) - \Pr_X(\hat{Y}_\tau = 1 \mid A^S = 1) \right| d\tau \\
&= \mathbb{E}_{\tau\sim U([0,1])} \Delta\mathrm{DP}^S(h_\tau),
\end{aligned}$$

and the last term is precisely the SPDD that we defined previously.

So far, we have observed bounding SPDD amounts control the 1-Wasserstein distance between distributions of score values. Our goal is to achieve fairness with respect to both A^S and A^T while retaining some theoretical control over the classification performance for the target dataset whose label information is obscured. Based on the idea that the reduction of Wasserstein distance is the key for fairness, we take advantage of neural network architecture with an intermediate (latent) layer. As we illustrate in the following, training the latent representation to match major and minor groups from the source and target distributions, one can establish bounds

on both the disparity and the accuracy of the classifier on the target dataset.

Let $\mathcal{Z} \subset \mathbb{R}^{d_\mathcal{Z}}$ be the latent space, and let $g_\phi : \mathcal{X} \to \mathcal{Z}$ be the "encoding map," parametrized by ϕ. Next, we consider a parametrized family of latent score functions $f_\theta : \mathcal{Z} \to [0, 1]$, and we assume that for all θ, $\|f_\theta\|_L \leq K$, for some fixed constant $K > 0$. This assumption is not artificial when we are working with neural network based score functions, since the usage of nonexpansive activations such as ReLU, together with spectral regularization on linear layers, provides adequate control on the Lipschitz constant. We write $\eta_{\phi,\theta}$ for the score function $f_\theta \circ g_\phi : \mathcal{X} \to [0, 1]$.

Next, denote the latent distributions induced by g_ϕ as

$$\nu_a^S = (g_\phi)_\# \mathcal{D}_a^S, \quad \nu_a^T = (g_\phi)_\# \mathcal{D}_a^T$$

for $a = 0, 1$. The score distributions, again induced from ν's by f_θ, are denoted as

$$\mu_a^S = (f_\theta)_\# \nu_a^S, \quad \mu_a^T = (f_\theta)_\# \nu_a^T.$$

Note that all the above distributions have dependency on parameters, but we have made them implicit in order to keep notations concise. Clearly, $W_1(\mu_0^S, \mu_1^S) = \mathrm{SPDD}^S(\eta_{\phi,\theta})$ and $W_1(\mu_0^T, \mu_1^T) = \mathrm{SPDD}^T(\eta_{\phi,\theta})$, as we have already shown. However, we step further and show that, in the following proposition, the deviation of SPDD^T from SPDD^S is bounded by the discrepancy of the source and target latent distributions, in terms of Wasserstein distance.

Proposition 1: Let $f_\theta : \mathcal{Z} \to [0, 1]$ be K-Lipschitz. Then,

$$\mathrm{SPDD}^T(\eta_{\phi,\theta}) \leq \mathrm{SPDD}^S(\eta_{\phi,\theta}) + K[W_1(\nu_0^S, \nu_0^T) + W_1(\nu_1^S, \nu_1^T)]. \quad (5)$$

Proof: Using Kantorovich-Rubinstein duality, we have

$$\begin{aligned} W_1(\mu_0^S, \mu_0^T) &= \sup_{\|u\|_L \leq 1, u:\Omega \to \mathbb{R}} \mathbb{E}_{\tau\sim\mu_0^S}[u(\tau)] - \mathbb{E}_{\tau\sim\mu_0^T}[u(\tau)] \\ &= \sup_{\|u\|_L \leq 1, u:\Omega \to \mathbb{R}} \mathbb{E}_{z\sim\nu_0^S}[u \circ f_\theta(z)] - \mathbb{E}_{z\sim\nu_0^T}[u \circ f_\theta(z)] \\ &\leq \sup_{\|f\|_L \leq K, f:\mathcal{Z} \to \mathbb{R}} \mathbb{E}_{z\sim\nu_0^S}[f(z)] - \mathbb{E}_{z\sim\nu_0^T}[f(z)] \\ &= K \cdot W_1(\nu_0^S, \nu_0^T), \end{aligned}$$

where in the third line we used that f_θ is K-Lipschitz and thus its composition with any 1-Lipschitz $u : \Omega \to \mathbb{R}$ is K-Lipschitz. Similarly,

$$W_1(\mu_1^S, \mu_1^T) \leq K \cdot W_1(\nu_1^S, \nu_1^T).$$

Therefore, using the triangular inequality for 1-Wasserstein distance,

$$\begin{aligned} &\mathrm{SPDD}^T(\eta_{\phi,\theta}) \\ &\quad = W_1(\mu_0^T, \mu_1^T) \\ &\quad \leq W_1(\mu_0^T, \mu_0^S) + W_1(\mu_0^S, \mu_1^S) + W_1(\mu_1^S, \mu_1^T) \\ &\quad \leq K \cdot W_1(\nu_0^T, \nu_0^S) + \mathrm{SPDD}^S(\eta_{\phi,\theta}) + K \cdot W_1(\nu_1^S, \nu_1^T) \\ &\quad = \mathrm{SPDD}^S(\eta_{\phi,\theta}) + K[W_1(\nu_0^S, \nu_0^T) + W_1(\nu_1^S, \nu_1^T)]. \end{aligned}$$

□

We would like to remark that using the distributions $\mathcal{D}_{a,y}^S$ doubly conditioned on $A^S = a$ and $Y^S = y$ with $a = 0, 1$, $y = 1$, one could derive similar results for SPDOp.

4) GENERALIZATION BOUND FOR DOMAIN TRANSFER

Proposition 1 motivates the parameter selection which jointly minimizes $W_1(\mu_0^S, \mu_1^S)$ and $W_1(\nu_a^S, \nu_a^T)$ for $a = 0, 1$. Indeed, this is what our framework does. (see Figure 2.) We further justify this methodology by providing a generalization bound that is analogous to the error bound in classical domain adaptation problems using the $\mathcal{H}$-divergence. Here, we generalize the approach of [30] and give a careful analysis suitable for our setting.

It is commonly assumed in the domain adaptation literature that there exists true labeling function $h^* : \mathcal{X} \to \{0, 1\}$, which is shared between the source and the target distributions and assigns a correct label to every sample. However, in our scenario, the label means either success or failure and there is usually no absolute sense of correctness on the assignment of the label. Hence we promote an alternative setup wherein we assume the existence of *true score function* $\eta^* : \mathcal{X} \to [0, 1]$, which assigns to each sample x the (true) probability of its acceptance (or success) $\eta^*(x)$. It soon becomes clear that this also makes the following analysis compatible with our choice of Wasserstein distance as a discrepancy measure for distributions.

As we preferred scores to labels,

Definition 2: The *relative risk* between the two score functions $\eta, \eta' : \mathcal{X} \to [0, 1]$ under the source distribution is defined as

$$\varepsilon^S(\eta, \eta') = \mathbb{E}_{X\sim\mathcal{D}^S} |\eta(X) - \eta'(X)|.$$

The relative risk for each sensitive group $A^S = a$ $(a = 0, 1)$ is similarly given by

$$\varepsilon_a^S(\eta, \eta') = \mathbb{E}_{X\sim\mathcal{D}_a^S} |\eta(X) - \eta'(X)|.$$

The *risk* of a score function η is then defined as

$$\varepsilon^S(\eta) = \varepsilon^S(\eta, \eta^*), \quad \varepsilon_a^S(\eta) = \varepsilon_a^S(\eta, \eta^*).$$

Analogous concepts for the target distribution could be defined in the same way, with superscripts S replaced by T.

To demonstrate that the definition is not a mere artifact for the discussion, we present the following proposition, which bridges our definition of risk and the traditional one based on labels.

Proposition 2: Let $\eta, \eta' : \mathcal{X} \to [0, 1]$ be score functions, and for each $\tau \in (0, 1)$, let h_τ, h'_τ be the corresponding classifiers with threshold τ, i.e.,

$$h_\tau(x) = 1_{\{\eta(x)>\tau\}}(x), \quad h'_\tau(x) = 1_{\{\eta'(x)>\tau\}}(x)$$

Then we have

$$\varepsilon(\eta, \eta') = \int_0^1 \Pr_{X\sim\mathcal{D}}(h_\tau(X) \neq h'_\tau(X))\, d\tau, \quad (6)$$

where $\mathcal{D}$ is a distribution of interest and ε is the associated relative risk.

Equation (6) can be paraphrased as: The relative risk measures the probability that h_τ and h'_τ disagree with their predictions, averaged over τ.

Proof: We begin with noting that

$$\begin{aligned}\Pr_{X\sim\mathcal{D}}(h_\tau(X)\neq h'_\tau(X)) &= \mathbb{E}_{X\sim\mathcal{D}}\left[1_{\{h_\tau(X)\neq h'_\tau(X)\}}(X)\right]\\ &= \mathbb{E}_{X\sim\mathcal{D}}\left|h_\tau(X)-h'_\tau(X)\right|.\end{aligned}$$

Hence

$$\begin{aligned}&\int_0^1 \Pr_{X\sim\mathcal{D}}(h_\tau(X)\neq h'_\tau(X))\,d\tau\\ &= \int_0^1\int_{\mathcal{X}}\left|h_\tau(x)-h'_\tau(x)\right|\,d\mathcal{D}(x)\,d\tau\\ &= \int_0^1\int_{\mathcal{X}}|1_{\{\eta(x)>\tau\}}(x)-1_{\{\eta'(x)>\tau\}}(x)|\,d\mathcal{D}(x)\,d\tau\\ &= \int_{\mathcal{X}}\int_0^1|1_{\{\eta(x)>\tau\}}(x)-1_{\{\eta'(x)>\tau\}}(x)|\,d\tau\,d\mathcal{D}(x)\\ &= \int_{\mathcal{X}}|\eta(x)-\eta'(x)|\,d\mathcal{D}(x)\\ &= \mathbb{E}_{X\sim\mathcal{D}}|\eta(X)-\eta'(X)| = \varepsilon(\eta,\eta'),\end{aligned}$$

where the third equality used Fubini's theorem and the fourth equality comes from

$$|1_{\{\eta(x)>\tau\}}(x)-1_{\{\eta'(x)>\tau\}}(x)| = \begin{cases}1 & \text{if } \tau\in(\eta(x),\eta'(x))\\ 0 & \text{otherwise.}\end{cases}$$

□

Let the latent encoding parameter ϕ be fixed, and let $\theta^* = \theta^{\text{opt}}(\phi)$ be the optimal parameter, in that it minimizes the combined risk

$$\lambda(\theta) = \varepsilon_0^S(\eta_{\phi,\theta}) + \varepsilon_1^S(\eta_{\phi,\theta}) + \varepsilon^T(\eta_{\phi,\theta}).$$

Let $\lambda^* = \lambda(\theta^*)$ be the optimal combined risk. We now state the generalization bound for our setting.

Theorem 1: Let ϕ be fixed. Let ν_a^S and ν_a^T ($a = 0, 1$) be corresponding latent distributions, and $\theta^* = \theta^{\text{opt}}(\phi)$ the optimal latent score parameter with minimal combined risk λ^*. If f_θ is K-Lipschitz for all θ, then the following holds for any value of θ:

$$\begin{aligned}\varepsilon^T(\eta_{\phi,\theta}) \le\ & \varepsilon_0^S(\eta_{\phi,\theta}) + \varepsilon_1^S(\eta_{\phi,\theta}) + \lambda^*\\ &+ 2K[W_1(\nu_0^S,\nu_0^T) + W_1(\nu_1^S,\nu_1^T)].\end{aligned}\tag{7}$$

Proof: By definition of ν's, for any θ, θ' and $a = 0, 1$ we have

$$\begin{aligned}\varepsilon_a^S(\eta_{\phi,\theta},\eta_{\phi,\theta'}) &= \mathbb{E}_{X\sim\mathcal{D}_a^S}|(f_\theta\circ g_\phi)(X)-(f_{\theta'}\circ g_\phi)(X)|\\ &= \mathbb{E}_{Z\sim\nu_a^S}|f_\theta(Z)-f_{\theta'}(Z)|\end{aligned}$$

and the same holds with superscripts T. Hence

$$\begin{aligned}&\varepsilon_a^T(\eta_{\phi,\theta},\eta_{\phi,\theta'}) - \varepsilon_a^S(\eta_{\phi,\theta},\eta_{\phi,\theta'})\\ &= \mathbb{E}_{Z\sim\nu_a^S}|f_\theta(Z)-f_{\theta'}(Z)| - \mathbb{E}_{Z\sim\nu_a^T}|f_\theta(Z)-f_{\theta'}(Z)|\\ &\le \sup_{\substack{f:\mathcal{Z}\to\mathbb{R}\\ \|f\|_L\le 2K}} \mathbb{E}_{Z\sim\nu_a^S}[f(Z)] - \mathbb{E}_{Z\sim\nu_a^T}[f(Z)]\\ &= 2K\cdot W_1(\nu_a^S,\nu_a^T),\end{aligned}$$

where the inequality holds since $|f_\theta - f_{\theta'}|$ is $2K$-Lipschitz. Therefore, we may proceed as

$$\begin{aligned}\varepsilon^T(\eta_{\phi,\theta},\eta_{\phi,\theta'}) &\le \varepsilon_0^T(\eta_{\phi,\theta},\eta_{\phi,\theta'}) + \varepsilon_1^T(\eta_{\phi,\theta},\eta_{\phi,\theta'})\\ &\le \varepsilon_0^S(\eta_{\phi,\theta},\eta_{\phi,\theta'}) + \varepsilon_1^S(\eta_{\phi,\theta},\eta_{\phi,\theta'})\\ &\quad + 2K[W_1(\nu_0^S,\nu_0^T) + W_1(\nu_1^S,\nu_1^T)].\end{aligned}$$

Now plugging in $\theta'\leftarrow\theta^*$ and combining with

$$\varepsilon^T(\eta_{\phi,\theta}) \le \varepsilon^T(\eta_{\phi,\theta^*}) + \varepsilon^T(\eta_{\phi,\theta},\eta_{\phi,\theta^*}),$$

we obtain

$$\begin{aligned}\varepsilon^T(\eta_{\phi,\theta}) &\le \varepsilon^T(\eta_{\phi,\theta^*}) + \varepsilon_0^S(\eta_{\phi,\theta},\eta_{\phi,\theta^*}) + \varepsilon_1^S(\eta_{\phi,\theta},\eta_{\phi,\theta^*})\\ &\quad + 2K[W_1(\nu_0^S,\nu_0^T) + W_1(\nu_1^S,\nu_1^T)]\\ &\le \varepsilon^T(\eta_{\phi,\theta^*}) + \varepsilon_0^S(\eta_{\phi,\theta^*}) + \varepsilon_1^S(\eta_{\phi,\theta^*})\\ &\quad + \varepsilon_0^S(\eta_{\phi,\theta}) + \varepsilon_1^S(\eta_{\phi,\theta})\\ &\quad + 2K[W_1(\nu_0^S,\nu_0^T) + W_1(\nu_1^S,\nu_1^T)]\\ &= \varepsilon_0^S(\eta_{\phi,\theta}) + \varepsilon_1^S(\eta_{\phi,\theta})\\ &\quad + 2K[W_1(\nu_0^S,\nu_0^T) + W_1(\nu_1^S,\nu_1^T)] + \lambda^*.\end{aligned}$$

□

The moral of Theorem 1 is that, reducing the quantity $W_1(\nu_0^S,\nu_0^T)+W_1(\nu_1^S,\nu_1^T)$ would result in a better upper bound for $\varepsilon^T(\eta_{\phi,\theta})$. Note that the quantity is also involved on the right-hand side of (5) in Proposition 1, which upper bounds the SPDD for the target group. Therefore, our framework, which aims to minimize both of the upper bounds in (5) and (7), is expected to achieve good model performance and fairness over the source and the target datasets.

B. PROPOSED METHOD

1) IMPLEMENTATION

We have training observations $(x_i^S, a_i^S, y_i^S)_{i=1}^{n_S}$ from the source domain and $(x_i^T, a_i^T, y_i^T)_{i=1}^{n_T}$ from the target domain (y_i^T are not provided in the unsupervised domain adaptation setting). We train the encoder g_ϕ and the latent score function f_θ with parameters ϕ and θ to be as fair and accurate as possible over both domains. Given an acceptance threshold τ, the prediction $(\hat{y}_\tau)_i$ for an input x_i is $1_{\{f_\theta(g_\phi(x))>\tau\}}(x_i)$.

a: SOURCE DOMAIN ACCURACY

To train the classifier to attain good prediction capability through supervised learning using the source data, we include the supervised cross-entropy loss

$$\mathcal{L}_{CE} = \frac{1}{n_S}\sum_{i=1}^{n_S} l(f_\theta(g_\phi(x_i^S)), y_i^S)),\tag{8}$$

where l denotes cross-entropy loss function.

b: SOURCE DOMAIN DEMOGRAPHIC PARITY (DP)

Since we should train a model with a score function that promotes demographic parity, we wish to include in our objective the 1-Wasserstein distance between score distributions between sensitive subgroups from the source domain.

However, the exact computation of the true score distributions is intractable, so we instead use empirical distributions $\hat{\mu}$, defined by

$$\hat{\mu}_a^S = \frac{1}{|B_a^S|} \sum_{i \in B_a^S} \delta_{f_\theta \circ g_\phi(x_i^S)} \tag{9}$$

for $a = 0, 1$, where δ_p is the Dirac measure centered at $p \in \mathbb{R}$, and B_a is a subset of the index set

$$I_a^S = \{i = 1, \ldots, n_S : a_i^S = a\}.$$

Then, we define the loss function

$$\mathcal{L}_{fair^S} = W_1(\hat{\mu}_0^S, \hat{\mu}_1^S). \tag{10}$$

The distributions in (9) are uniform mixtures of delta distributions centered at the sets B_a^S of samples drawn from the respective underlying distributions ($a = 0, 1$). Despite its simplicity, estimating empirical distributions in this manner is conducive to calculating empirical versions of 1-Wasserstein distance, because W_1 distance can be computed exactly by a simple closed-form expression for one-dimensional empirical distributions with equal number of point masses [32], as we present below. Note that the minimization of (10) can be viewed as a stochastic minimization of $\text{SPDD}^S(\eta_{\phi,\theta})$, which appears on the right-hand side of (5) in Proposition 1.

Let $\hat{\mu}_0 = \frac{1}{m}\sum_{i=1}^m \delta_{p_i}$ and $\hat{\mu}_1 = \frac{1}{m}\sum_{i=1}^m \delta_{q_i}$, where p_i, $q_i \in \mathbb{R}$ for $i = 1, \ldots, m$. Let $\rho : \mathbb{R}^m \to \mathbb{R}^m$ be a sorting function such that given a vector $\mathbf{r} = (r_1, \cdots, r_m)$, outputs

$$\rho(\mathbf{r}) = (r_{\sigma(1)}, \ldots, r_{\sigma(m)})$$

where $\sigma : \{1, \ldots, m\} \to \{1, \ldots, m\}$ is a rearrangement of indices such that $1 \le i < j \le m$ implies $r_{\sigma(i)} \le r_{\sigma(j)}$. Then the optimal coupling γ^* is simply given as the assignment $\rho(\mathbf{p})_i \mapsto \rho(\mathbf{q})_i$ for each $i = 1, \ldots, m$. That is, the 1-Wasserstein distance between $\hat{\mu}_0$ and $\hat{\mu}_1$ is given by

$$W_1(\hat{\mu}_0, \hat{\mu}_1) = \frac{1}{m}\sum_{i=1}^m |\rho(\mathbf{p})_i - \rho(\mathbf{q})_i|, \tag{11}$$

where m is the size of the mini-batch. Hereafter, we minimize the 1-Wasserstein distance between two one-dimensional distributions using equation (11).

Note that the number of point masses should be the same for the distributions to apply to this formulation. In our implementation, this condition will always be satisfied, since our choice of B_0^S and B_1^S are batches for stochastic gradient descent, and the batch size is a fixed constant m for all distributions under consideration.

c: DOMAIN ADAPTATION LOSS FOR DP

Taking as objective the weighted sum of the loss functions $\mathcal{L}_{CE}$ and $\mathcal{L}_{fair^S}$ we have introduced so far, and optimizing it with respect to parameters ϕ and θ via gradient descent, we obtain a classification model $f_\theta(g_\phi(x))$ that is accurate and fair with respect to the source domain A^S. To enforce the model to attain predictive power and fairness in the target domain, we introduce additional loss functions.

What we want to minimize is the 1-Wasserstein distance between the group-wise latent distributions on $\mathcal{Z}$ induced by the source and target domains. As for $\mathcal{L}_{fair^S}$, we use the empirical distributions

$$\hat{\nu}_a^S = \frac{1}{|B_a^S|} \sum_{i \in B_a^S} \delta_{g_\phi(x_i^S)}, \quad a = 0, 1,$$

where $B_a^S \subset I_a^S$, and δ_z is the Dirac measure centered at $z \in \mathcal{Z}$. Empirical distributions $\hat{\nu}_a^T$ for the target domain are defined similarly with superscripts S replaced by T.

For multi-dimensional empirical probability measures, however, calculating the Wasserstein distance is, in general, computationally costly. It requires finding an optimal coupling in (3), and for discrete distributions, this problem is equivalent to solving a linear program with number of variables proportional to the square of the number of samples. To circumvent this problem and exploit the efficient formulation (11), we adopt sliced Wasserstein discrepancy (SWD), which has been utilized as an approximation to Wasserstein distance in various works of machine learning literature [29], [32].

Sliced Wasserstein discrepancy is computed by first making random projections onto one-dimensional spaces, and then calculating W_1 distances between the projected measures. More precisely,

$$\text{SWD}(\hat{\nu}_0, \hat{\nu}_1) = \int_{\mathcal{S}} W_1(\hat{\nu}_0^w, \hat{\nu}_1^w)\, d\omega(w), \tag{12}$$

where ω is a uniform measure on the unit sphere $\mathcal{S} = \mathcal{S}^{d_Z - 1}$ in $\mathbb{R}^{d_Z}$ such that $\int_{\mathcal{S}} d\omega = 1$, and the measures $\hat{\nu}_a^w = w^\intercal \hat{\nu}_a$ are one-dimensional projections of $\hat{\nu}_a$ onto the direction of $w \in \mathcal{S}$. In practice, the intractable integration over $\mathcal{S}$ is usually replaced by an approximation through sampling.

Suppose that $\hat{\nu}_a$ has mass on $\{z_{1,a}, \ldots, z_{m,a}\} \subset \mathcal{Z}$, and denote $\mathbf{z}_a = (z_{1,a}, \ldots, z_{m,a})$ for $a = 0, 1$. Then the SWD can be approximated as

$$\begin{aligned}\text{SWD}(\hat{\nu}_0, \hat{\nu}_1) &\approx \frac{1}{|\hat{\mathcal{S}}|} \sum_{w \in \hat{\mathcal{S}}} W_1(\hat{\nu}_0^w, \hat{\nu}_1^w) \\ &= \frac{1}{|\hat{\mathcal{S}}|} \sum_{w \in \hat{\mathcal{S}}} \frac{1}{m} \sum_{i=1}^m |\rho(w^\intercal \mathbf{z}_0)_i - \rho(w^\intercal \mathbf{z}_1)_i|,\end{aligned} \tag{13}$$

where $\hat{\mathcal{S}} = \{w_j\}_{j=1}^k$ consists of k uniform samples from $\mathcal{S}$ and ρ is the sorting function mentioned previously. Using SWD, we replace the highly inefficient computation of W_1 by k one-dimensional optimal transport problems using (11), resulting in a computationally efficient algorithm.

Now we define our domain adaptation loss function on the minority group ($A^S = 0, A^T = 0$) as:

$$\mathcal{L}_{DA_0} = \text{SWD}(\hat{\nu}_0^S, \hat{\nu}_0^T).$$

and similarly for majority subgroup ($A^S = 1, A^T = 1$):

$$\mathcal{L}_{DA_1} = \text{SWD}(\hat{\nu}_1^S, \hat{\nu}_1^T).$$

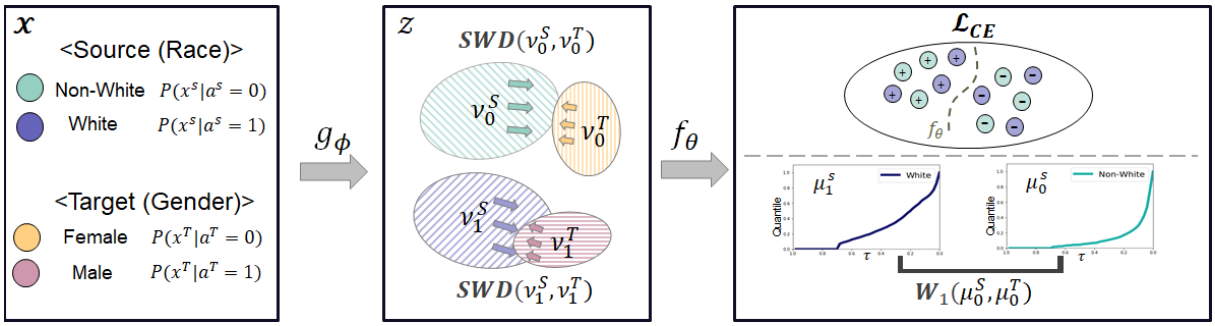


FIGURE 2. Illustration of the overall architecture of the proposed method for fairness criterion DP.

Note that the distributions $\hat{\nu}_0^T$ and $\hat{\nu}_1^T$ are distinguished according to sensitive attributes A^T in the target domain. Access to the label information from the target domain is not required, and our algorithm remains unsupervised.

Summing up, we arrive at our total loss function, on which we optimize the parameters ϕ and θ via stochastic gradient descent:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{fair^S} + \lambda_2(\mathcal{L}_{DA_0} + \mathcal{L}_{DA_1}), \qquad (14)$$

where λ_1 and λ_2 are hyperparameters to be tuned to balance the terms. The first and the third terms in (14) consist of the upper bound of the target risk in Theorem 1 (except the irreducible term λ^*), while the second and the third terms bound SPDD for the target domain, as shown in Proposition 1. Therefore, the proposed framework attempts to achieve model fairness as well as prediction performance, over both domains.

The architecture of our methodology is illustrated in Figure 2. This figure shows an example of our framework in which the sensitive attribute of the source domain is *race* (white/non-white) and that of the target domain is *gender* (female/male).

In the latent space $\mathcal{Z}$, by minimizing $\text{SWD}(\hat{\nu}_0^S, \hat{\nu}_0^T)$, we align the (empirical) distributions of minority groups "non-white" and "female", as shown in the middle of the figure. Similarly, the minimization of $\text{SWD}(\hat{\nu}_0^S, \hat{\nu}_0^T)$ aligns the latent distributions of the majority groups ("white" and "male"). The effect of minimizing the loss functions $\mathcal{L}_{CE}$ and $W_1(\hat{\mu}_0^S, \hat{\mu}_1^S)$ is shown in the rightmost figure. The supervised loss function $\mathcal{L}_{CE}$ guides the model to maximize the accuracy of its predictions, and the source domain DP loss $W_1(\hat{\mu}_0^S, \hat{\mu}_1^S)$ enforces the cumulative distribution functions for the score values of "non-white" and "white" groups to become similar.

d: EQUAL OPPORTUNITIES (EOP)

When the fairness criterion of interest is EOp, we need access to at least some of the labels from the target domain, because in this case, distributions of the form

$$\mathcal{D}_{a,1}^T = \mathcal{L}(X^T \,|\, A^T = a, Y^T = 1)$$

are involved in the definition of SPDOp. Therefore, in the following, we assume a semi-supervised domain adaptation setting in which we are partially provided with label information from the target domain dataset. Other than this point, one can carry out the analysis in the same way as one can for DP.

To achieve EOp in the target dataset, we need to reduce the Wasserstein distance between group-wise score distributions from the source domain, conditioned on $Y^S = 1$. Hence we consider the corresponding empirical distributions:

$$\hat{\mu}_{a,1}^S = \frac{1}{|B_{a,1}^S|} \sum_{i \in B_{a,1}^S} \delta_{f_\theta \circ g_\phi(x_i^S)}$$

where δ denotes the Dirac measure and $B_{a,y}^S$ are subsets of

$$I_{a,y}^S = \{i = 1, \ldots, n_S : a_i^S = a, y_i = y\}$$

for $a = 0, 1, y = 1$. The source domain fairness loss $\mathcal{L}_{fair^S}$ in terms of EOp then become:

$$\mathcal{L}_{fair^S} = W_1(\hat{\mu}_{0,1}^S, \hat{\mu}_{1,1}^S).$$

Similarly, the domain adaptation loss involve empirical latent distributions conditioned on $Y = 1$:

$$\hat{\nu}_{a,1}^S = \frac{1}{|B_{a,1}^S|} \sum_{i \in B_{a,1}^S} \delta_{g_\phi(x_i^S)},$$

$$\hat{\nu}_{a,1}^T = \frac{1}{|I_{a,1}^T|} \sum_{i \in B_{a,1}^T} \delta_{g_\phi(x_i^S)}.$$

With these notations, we define

$$\mathcal{L}_{DA_a} = \text{SWD}(\hat{\nu}_{a,1}^S, \hat{\nu}_{a,1}^T),$$

for $a = 0, 1$.

Finally, the total loss function $\mathcal{L}_{total}$ is given by the same formula (14) as in the DP case.

2) ALGORITHM

Algorithm 1 outlines the detailed procedure for the proposed method when the fairness criterion is DP. Parameters are optimized stochastically through batch-wise gradient descent. Implementation is done by using the deep learning library *PyTorch*.

Algorithm 1 Batch Training Procedure for the Proposed Method (DP)

Input: Labeled source dataset $(x_\ell^S, a_\ell^S, y_\ell^S)_{\ell=1}^{n_S} \sim \mathcal{D}^S$, unlabeled target dataset $(x_\ell^T, a_\ell^T)_{\ell=1}^{n_T} \sim \mathcal{D}^T$, and a randomly initialized encoder g_ϕ and latent score function f_θ, hyperparameters λ_1, λ_2, number of projections k, batch size m, latent dimension $d_\mathcal{Z}$, and learning rate α

Output: Trained encoder g_ϕ and trained latent score function f_θ.

for *Mini-batch* B_0^S,B_1^S *from source dataset, mini-batch* B_0^T,B_1^T *from target dataset, where* $|B_0^S| = |B_1^S| = |B_0^T| = |B_1^T| = m$ **do**

- Calculate domain transfer SWD loss with minority subgroup (B_0^S,B_0^T) as:
  - Obtain representations in latent space $g_\phi(B_0^S)$ and $g_\phi(B_0^T)$ from mini-batch from minority subgroup;
  - Sample $\{w_j\}_{j=1}^k$ from the unit sphere $\mathcal{S}^{d_\mathcal{Z}-1}$;
  - Sort $w_j^\intercal g_\phi(B_0^S)$ with sorting algorithm ρ such that $\rho(w_j^\intercal g_\phi(x))_i \le \rho(w_j^\intercal g_\phi(x))_{i+1},\ \forall x \in B_0^S,\ 1 \le i < m$;
  - Sort $w_j^\intercal g_\phi(B_0^T)$ with sorting algorithm ρ such that $\rho(w_j^\intercal g_\phi(x))_i \le \rho(w_j^\intercal g_\phi(x))_{i+1},\ \forall x \in B_0^T,\ 1 \le i < m$;
  - $\mathcal{L}_{DA_0} = \frac{1}{m}\frac{1}{k}\sum_{i=1}^m \sum_{j=1}^k |\rho(w_j^\intercal g_\phi(B_0^S))_i - \rho(w_j^\intercal g_\phi(B_0^T))_i|$;
- Calculate domain transfer SWD loss with majority subgroup (B_1^S,B_1^T) as:
  - Obtain representations in latent space $g_\phi(B_1^S)$ and $g_\phi(B_1^T)$ from mini-batch from majority subgroup;
  - Sample $\{w_j\}_{j=1}^k$ from the unit sphere $\mathcal{S}^{d_\mathcal{Z}-1}$;
  - Sort $w_j^\intercal g_\phi(B_1^S)$ with sorting algorithm ρ such that $\rho(w_j^\intercal g_\phi(\tilde{x}))_i \le \rho(w_j^\intercal g_\phi(\tilde{x}))_{i+1},\ \forall \tilde{x} \in B_1^S,\ 1 \le i < m$;
  - Sort $w_j^\intercal g_\phi(B_1^T)$ with sorting algorithm ρ such that $\rho(w_j^\intercal g_\phi(\tilde{x}))_i \le \rho(w_j^\intercal g_\phi(\tilde{x}))_{i+1},\ \forall \tilde{x} \in B_1^T,\ 1 \le i < m$;
  - $\mathcal{L}_{DA_1} = \frac{1}{m}\frac{1}{k}\sum_{i=1}^m \sum_{j=1}^k |\rho(w_j^\intercal g_\phi(B_1^S))_i - \rho(w_j^\intercal g_\phi(B_1^T))_i|$;
- Calculate W1 loss with source data set(B_0^S,B_1^S) as:
  - Obtain classifier output $f_\theta(g_\phi(B_0^S))$ and $f_\theta(g_\phi(B_1^S))$ from mini-batch from source data set;
  - Sort $f_\theta(g_\phi(B_0^S))$ with sorting algorithm ρ such that $\rho(f_\theta(g_\phi(x)))_i \le \rho(f_\theta(g_\phi(x)))_{i+1},\ \forall x \in B_0^S,\ 1 \le i < m$;
  - Sort $f_\theta(g_\phi(B_1^S))$ with sorting algorithm ρ such that $\rho(f_\theta(g_\phi(\tilde{x})))_i \le \rho(f_\theta(g_\phi(x)))_{i+1},\ \forall \tilde{x} \in B_1^S,\ 1 \le i < m$;
  - $\mathcal{L}_{fair^S} = \frac{1}{m}\sum_{i=1}^m |\rho(f_\theta \circ g_\phi(B_0^S))_i - \rho(f_\theta \circ g_\phi(B_1^S))_i|$;
- Calculate Supervised loss ;
  - $\mathcal{L}_{CE} = \sum_{(x_i^S, y_i^S) \in B_0^S} l(f_\theta(g_\phi(x_i^S)), y_i^S) + \sum_{(\tilde{x}_i^S, \tilde{y}_i^S) \in B_1^S} l(f_\theta(g_\phi(\tilde{x}_i^S)), \tilde{y}_i^S)$;
- Calculate total loss;
  - $\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{fair^S} + \lambda_2(\mathcal{L}_{DA_0} + \mathcal{L}_{DA_1})$;
- Update ϕ and θ with gradient descent;
  - $\phi \leftarrow \phi - \alpha \nabla_\phi \mathcal{L}_{total}$;
  - $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{total}$;

end

In Algorithm 1, we denote the mini-batch of data from the minority group $A^S = 0$ of the source dataset as B_0^S and that of the majority group $A^S = 1$ as B_1^S. The counterparts for the target dataset are denoted as B_0^T and B_1^T, respectively. Size of all mini-batches are set to a fixed integer m. The algorithm for EOp is basically the same, except that the batches are drawn among samples with $Y = 1$.

IV. EXPERIMENTS

In this section, we present how fairness was transferred over domains via our proposed algorithm for three real datasets. All datasets used contained two distinct sensitive attributes, and we experimented two-ways for each dataset, switching the roles of source and target. For example, when two sensitive attributes of a dataset were *race* and *gender*, we first tested the case where the attribute *race* belongs to the source domain and the *gender* to the target domain. Then, we performed the test again with the opposite scenario, i.e., where the sensitive attribute of the source data is *gender* and that of the target data is *race*.

A. EXPERIMENTAL SETTINGS

1) DATASETS

Adult: The UCI Adult dataset contains census information of 32565 individuals describing adults from the 1994 US Census. The classification goal of this Adult dataset is to predict whether an individual's annual income exceeds \$50K/year. All categorical variables were converted into multiple binary variables using one-hot encoding. The numerical variables, *age* and *workhours* were minimax normalized to exhibit values in [0,1]. The resulting preprocessed dataset had 15 input variables. We used the following choice of sensitive variables, to measure the fairness metric and to split the source and the target datasets: *race* [non-white ($A = 0$) / white ($A = 1$)] and *gender* [female ($A = 0$) / male ($A = 1$)].

COMPAS: The ProPublica COMPAS dataset contains 6172 instances that predict whether a criminal defendant will

recidivate within two years. All of the categorical variables were one-hot encoded and the numerical variable *# of priors* was minimax normalized to exhibit values in [0,1]. The resulting preprocessed dataset had 13 input variables. For sensitive attributes, we considered *race* [non-white ($A = 0$) / white ($A = 1$)] and *gender* [female ($A = 0$) / male ($A = 1$)].

German: The UCI German credit data contain 1000 data samples and attempts to predict each individual's credit risk (good/bad). All of the categorical variables were one-hot encoded and all of the numerical variables *duration*, *amount*, *installment*, *present resid*, *# of credits*, and *# of people* were minimax normalized. As a result, we used 56 input variables. For sensitive attributes, we used *age* [young (≤ 30) ($A = 0$) / old (≥ 30) ($A = 1$)] and *gender* [female ($A = 0$) / male ($A = 1$)].

2) COMPARATIVE MODELS

In this paper, we compared our method with following three models.

Baseline: Basic linear neural network classifier with no fairness constraints, trained using the loss function $\mathcal{L}_{CE}$ only.

W1(Source): Neural network classifier with fairness constraints on the source data only, trained using the loss functions $\mathcal{L}_{CE}$ and $\mathcal{L}_{fair^S}$.

MMD [24]: Domain adaptive fairness approach that tries to minimize discrepancy of fairness metrics between source and target domains by applying MMD loss in the latent space.

3) TRAINING DETAILS

All of the experimented models (three comparison models and ours) in this paper share the same neural network architecture. The encoding map g_ϕ is composed of two fully connected layers with ReLU activation functions FC($d \to d/2$) - ReLU - FC($d/2 \to d/4$) - ReLU, where d is the input space dimension. The latent score function f_θ has the neural network structure FC($d/4 \to 2$) - ReLU - FC($2 \to 1$) - Sigmoid.

All datasets were divided as follows: 70% training set, 10% validation set, and 20% test set. As explained above, our framework enables one to select a fairness criterion to be satisfied by taking appropriate conditional distributions. In our work, we selected and experimented with two of the most widely used fairness criteria: DP and EOp.

In the case of DP, our method does not require any information regarding the true label; therefore, we worked on a completely unsupervised domain adaptation setting. We simply split the training set to obtain equally sized source and target datasets. The source data instances contained both the inputs X and the label Y, while the target data instances contained only the input X. (Hence, the target labels were not used in model training.)

However, to achieve EOp, label information for the target data was partially required because we had to match the cumulative distribution of class and label-conditional scores. Thus, in this case, we assumed a semi-supervised domain adaptation setting, in which 20% of the target data contained label information.

We used Xavier initialization to initialize the weight with mini batch size $m = 128$ for Adult and COMPAS, and $m = 32$ for German. We optimized the network with ADAM [33] optimizer with learning rate $\alpha = 0.001$ for 100 epochs. The sampling number $k = 10$ was used for approximating SWD. To control the Lipschitz constant of score functions, we used L_2 spectral regularizers designed for network weight decay. We used the validation set to select values of λ_1, λ_2.

4) METRIC

To evaluate the models based on experimental results, we introduce metrics that measure the domain transfer performance of the model in terms of fairness. When we are interested in attaining demographic parity, we want both the source and the target values of SPDD to be close to 0. Hence, we take the maximum value of SPDD over domains,

$$\mathrm{SPDD}_{\max} = \max\{\mathrm{SPDD}^S(\eta_{\phi,\theta}), \mathrm{SPDD}^T(\eta_{\phi,\theta})\},$$

which can be viewed as the worst-case of fairness in terms of SPDD, as our measure of model performance for fairness in the DP setting.

In the same way, in the EOp setting, we define and use:

$$\mathrm{SPDOp}_{\max} = \max\{\mathrm{SPDOp}^S(\eta_{\phi,\theta}), \mathrm{SPDOp}^T(\eta_{\phi,\theta})\}.$$

B. RESULTS

Table 1 shows the experimental results for the transfer of DP on the three datasets. Values of the accuracy and $\mathrm{SPDD}_{\max}$ on the test data, averaged over 10 repeated experiments, are shown. Here, superscript * indicates a statistically significant difference between the results of our method and **MMD**, according to Wilcoxon signed rank test with level of significance 0.05. As shown, our method more effectively reduced $\mathrm{SPDD}_{\max}$ in both the source and the target domains, as compared to other models. We observed that **W1(Source)**, in general, could not reduce SPDD in both domains, except for the case of *gender* $\to$ *race* in the Adult dataset and *age* $\to$ *gender* in the German dataset. Additionally, the **MMD** often did transfer fairness from the source domain to the target domain, but **Ours** showed the lowest value of $\mathrm{SPDD}_{\max}$ in five out of six experiments, and four were significantly better than **MMD**.

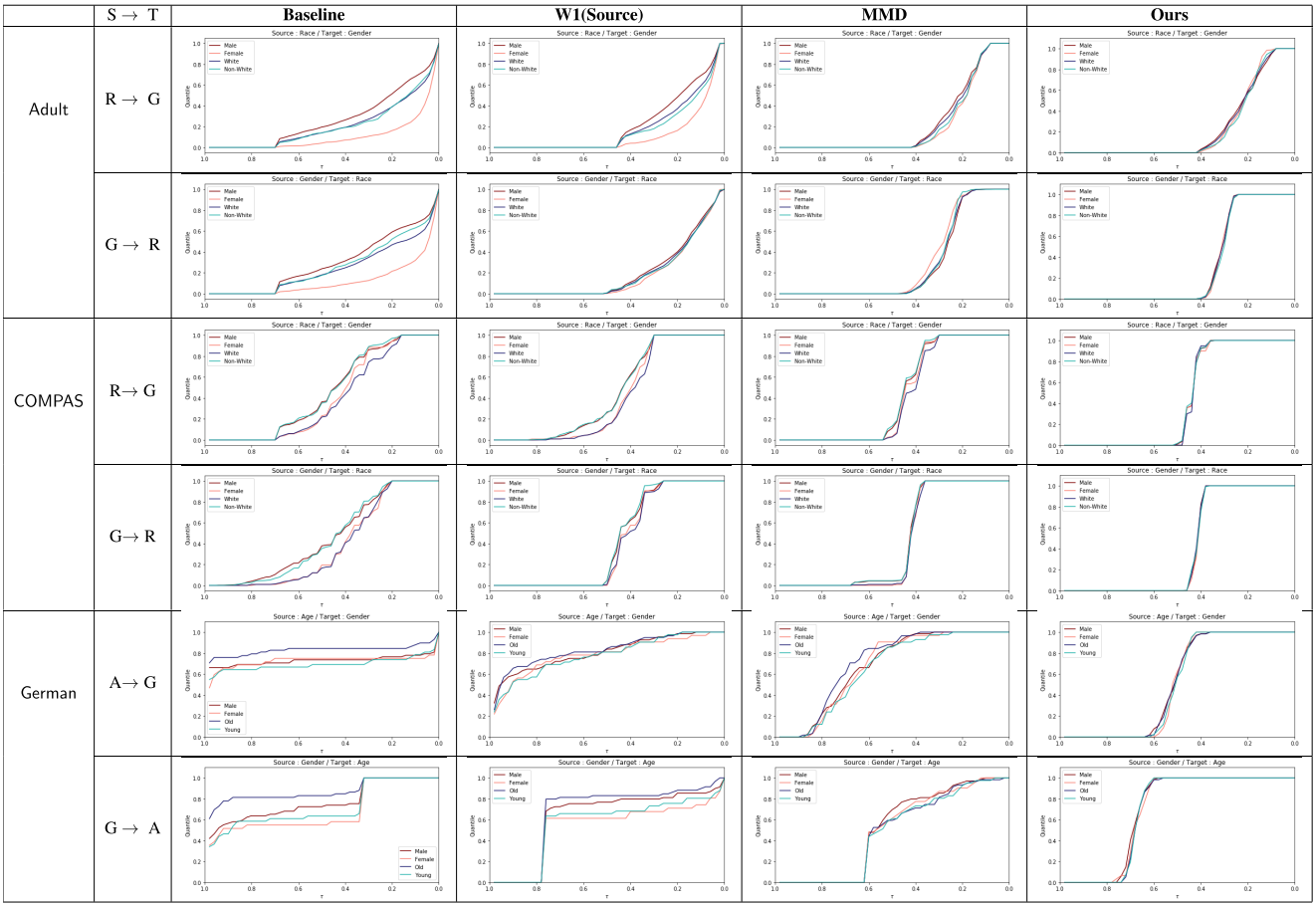
Due to the aforementioned intrinsic trade-off between prediction performance and fairness, in models with fairness consideration, accuracy degradation compared to the **Baseline** was observed. However, note that, while the other methods failed to meet the fairness criterion in some tasks, ours successfully achieved fairness in all of them. That is, our method best repaid the accuracy loss.

Table 2 shows the cumulative distribution of scores in the source and target domains. For the Adult and COMPAS datasets, red: male; orange: female; navy: white; and green: non-white. For the German dataset, red: male; orange:

TABLE 1. Experimental results comparing the suggested model and comparative models in an unsupervised domain adaptation setting. We aim to jointly minimize the demographic parity in the source and the target data. Values of models' test set accuracy and SPDD_{max} are shown. The numbers in bold indicate the best performance.

Dataset	Source → Target	Metric	Baseline	W1(Source)	MMD	Ours
Adult	*race* → *gender*	Accuracy	**0.8431**	0.8377	0.8023	0.8092
		SPDD_{max}	0.2098	0.2166	0.0781	**0.0398***
	gender → *race*	Accuracy	**0.8471**	0.8203	0.7924	0.8043
		SPDD_{max}	0.2606	0.0389	0.0721	**0.0248***
COMPAS	*race* → *gender*	Accuracy	**0.6576**	0.6401	0.6026	0.5874
		SPDD_{max}	0.1291	0.0927	0.1147	**0.0385***
	gender → *race*	Accuracy	**0.6830**	0.6506	0.5916	0.6030
		SPDD_{max}	0.1029	0.0967	0.0407	**0.0389**
German	*age* → Gender	Accuracy	**0.8715**	0.8066	0.8390	0.7997
		SPDD_{max}	0.1827	0.0346	0.0665	**0.0227***
	gender → *age*	Accuracy	**0.8807**	0.8613	0.7364	0.8046
		SPDD_{max}	0.2290	0.1337	**0.0378**	0.0412

TABLE 2. Experimental results that show cumulative distributions of score predictions μ_0, μ_1 in source and target domains. Each domains is denoted in capital letters, as "R": *race*,"G": *gender*, and "A": *age*. The red and orange lines indicate distributions for major and minor groups in one domain, and navy and green lines denote the major and minor distributions in the other domain.(The closer the four lines are, the better.)



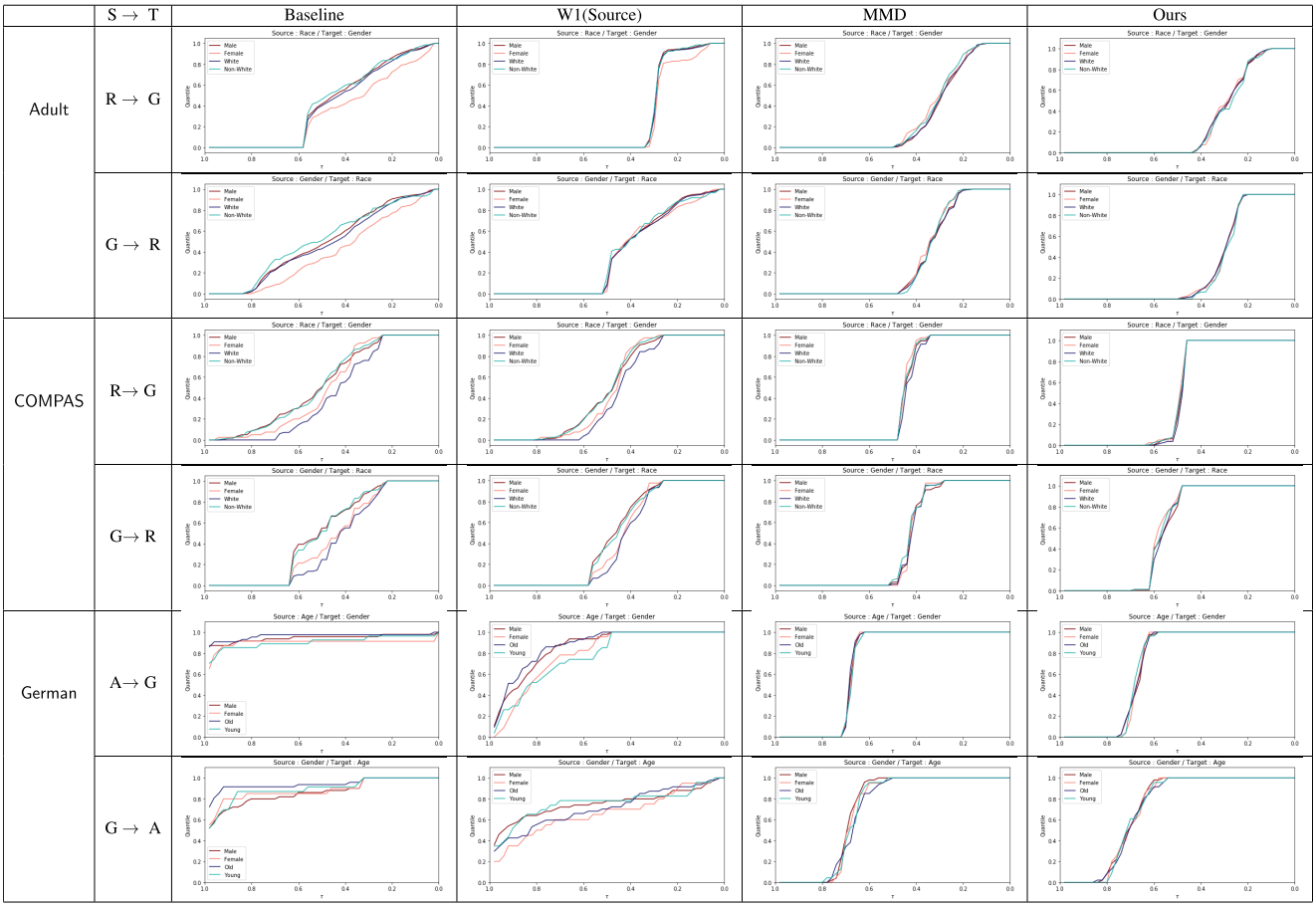
female; navy: white; and green: non-white. The difference between the red and orange lines indicates the SPDD in one domain, whereas that between the navy and green lines indicates the SPDD in the other domain. In this figure, the closer the four lines are to each other, the fairer is the result.

In the results for the Adult dataset with **Baseline** applied, the navy and green lines are already close to each other, but the red and orange lines are distant from each other. Thus, in this task, the DP gap for *race* is small, but that for *gender* is large. Therefore, the results from **W1(source)** for the task

TABLE 3. Experimental results comparing the suggested model and comparative models in semi-supervised domain adaptation setting. We aim to jointly minimize the Equalized Opportunity (EOp) gaps in the source and the target data. Values of models' test set accuracy and $\text{SPDOp}_{\text{max}}$ are shown. The numbers in bold indicate the best performance.

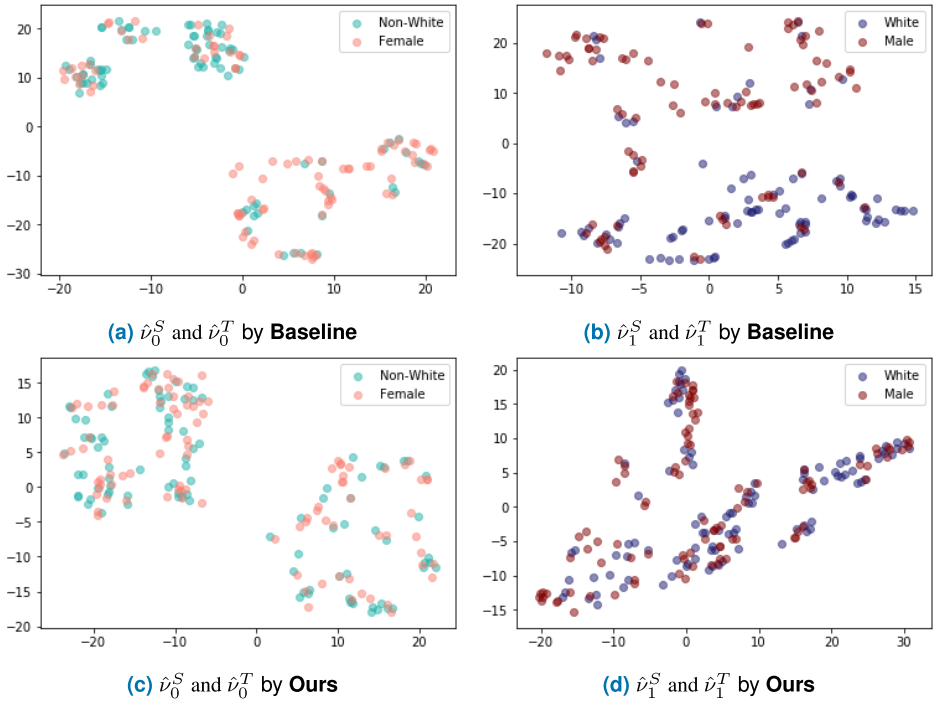
Dataset	Source → Target	Metric	Baseline	W1(Source)	MMD	Ours
Adult	*race* → *gender*	Accuracy	**0.8431**	0.8095	0.7984	0.7994
		$\text{SPDOp}_{\text{max}}$	0.1220	0.1009	0.0444	**0.0272***
	gender → *race*	Accuracy	**0.8471**	0.8381	0.7924	0.8014
		$\text{SPDOp}_{\text{max}}$	0.1150	0.0311	0.0330	**0.0301**
COMPAS	*race* → *gender*	Accuracy	**0.6576**	0.6486	0.6025	0.6043
		$\text{SPDOp}_{\text{max}}$	0.1324	0.1054	0.0925	**0.0446***
	gender → *race*	Accuracy	**0.6830**	0.6586	0.5986	0.5940
		$\text{SPDOp}_{\text{max}}$	0.1953	0.1293	**0.0395**	0.0427
German	*age* → *gender*	Accuracy	**0.8715**	0.7795	0.7007	0.7064
		$\text{SPDOp}_{\text{max}}$	0.1347	0.2015	**0.0381**	0.0521
	gender → *age*	Accuracy	**0.8807**	0.7821	0.6848	0.7944
		$\text{SPDOp}_{\text{max}}$	0.0914	0.1167	0.0605	**0.0317***

TABLE 4. Experimental results that show cumulative distributions of score predictions $\mu_{0,1}, \mu_{1,1}$ in source and target domains. Each domains were denoted in capital letters, as "R": *race*,"G": *gender*, and "A": *age*. The red and orange lines indicate distributions for major and minor groups in one domain, and navy and green lines denote the major and minor distributions in the other domain.(The closer the four lines are, the better.)



gender → *race* were successful, with SPDD_{max} reduced. However, it failed to achieve fairness in the reverse task *race* → *gender*. On the other hand, all four lines for **Ours** were made close to each other, in both domain adaptation settings. This implies that our method was capable of making fair decisions in all given settings.

Results for **Baseline** and **W1(source)** models on COMPAS and German data show that it would be more



(a) $\hat{\nu}_0^S$ and $\hat{\nu}_0^T$ by **Baseline** (b) $\hat{\nu}_1^S$ and $\hat{\nu}_1^T$ by **Baseline**

(c) $\hat{\nu}_0^S$ and $\hat{\nu}_0^T$ by **Ours** (d) $\hat{\nu}_1^S$ and $\hat{\nu}_1^T$ by **Ours**

FIGURE 3. t-SNE visualization of latent representations from **Baseline** and **Ours**.

challenging to train a jointly fair model for these datasets, because no pair of lines for any sensitive attribute is aligned. Even **MMD** was not very successful, except in the *gender* → *race* task on the COMPAS dataset. However, the proposed method consistently showed superior results in terms of fairness, as illustrated by the tightly bound four lines in the column for our method.

The experimental results for EOp are summarized in Table 3, which shows the average accuracy and $\text{SPDOp}_{\max}$ on the test data over 10 repeated experiments. The proposed method showed the best transfer of fairness in four out of six experimental settings, and three of them were significantly better than the comparative model **MMD**. **W1(source)** demonstrated good results on the task *gender* → *race* for the Adult dataset, but failed to achieve joint fairness on other situations.

Cumulative distributions of predicted scores for EOp are shown in Table 4. Similar to the case of DP, in this experiment, results from the **Baseline** and **W1(source)** models showed gaps in the four lines representing the cumulative distributions. On the other hand, **Ours** demonstrated satisfactory results, with the four lines all aligned.

In standard transfer learning problems, negative transfer might occur [34]. It is a phenomenon in which transfer learning degrades the performance on the target domain, instead of improving it, usually when the source domain is irrelevant to the target domain. However, in our task, the only difference between the domains is that the source and the target domain have distinct sensitive attributes. Therefore, negative transfer is less likely to occur in our settings, since the source and the target domains are closely related to each other. The experimental results verified that our method was indeed effective in improving the target domain disparity compared to **Baseline** and **W1(Source)**.

C. VISUALIZATION

The domain adaptive loss function $\mathcal{L}_{DA_0}$ is designed for learning latent representations in which the distribution of two minority groups (non-white and female) is similar. In the same way, the latent distribution of majority groups (white and male) should become alike in the presence of $\mathcal{L}_{DA_1}$. To verify that the suggested model is as well trained as we intended, we visualized the latent embedding of the source and the target data using t-SNE [35].

Figure 3 shows the visualization of $\hat{\nu}_0^S$, $\hat{\nu}_0^T$, $\hat{\nu}_1^S$, and $\hat{\nu}_0^T$ from the Adult dataset, produced by **Baseline** and **Ours** with $\text{SPDD}_{\max}$ as fairness criterion. While the latent distributions from the two domains tended be separated when the baseline method was used, our method successfully aligned the distribution of minor groups and major groups. For example, in the right figures, embedded distributions of the red (male) and blue (white) samples seem to be concentrated on distinct clusters (Figure 3c), whereas, in the latent space of our model, the distributions are more evenly blended (Figure 3d).

V. DISCUSSION

In this study, we solved a fair classification problem, wherein the sensitive attributes of the data used for model learning differed from the sensitive attributes of the test data. We applied the concept of domain adaptation for fair machine learning to

address the problem. We proposed a new method that learns similar latent representations for the source and the target domain minority groups via introducing sliced Wasserstein discrepancy loss function, and for the majority groups as well. Simultaneously, 1-Wasserstein distance between the predicted score distributions of the minority and the majority groups in the source domain was also minimized. Adapting ideas from the domain adaptation literature, we derived a generalization bound that provided control over the classification performance of the trained model in two domains. Consequently, we arrived at a method that achieves joint fairness for different sensitive attributes, with some flexibility over the choice of fairness criteria (DP and EOp). Notably, the classification fairness of our model was robust to the choice of threshold on predicted score values, as our method aligns the cumulative score distributions of each demographic subgroup.

We empirically tested our method on three standard datasets: Adult, COMPAS, and German. Comparison of $\text{SPDD}_{\max}$ showed that our method outperformed previous methods in terms of mutual fairness and, as we had predicted by the theory, the accuracy degradation was within a reasonable range. Additionally, by visualizing latent representations, we demonstrated that using Wasserstein-based loss was indeed effective in blending latent distributions.

We believe that our method could be extended to more intricate tasks pertaining to domain adaptation and fair machine learning. We plan to continue pursuing the study of model fairness, possibly by incorporating more challenging domain adaptation settings. We also anticipate that our method could be generalized for different notions of fairness, including individual fairness.

REFERENCES

[1] M. Yurochkin, A. Bower, and Y. Sun, "Training individually fair ML models with sensitive subspace robustness," in *Proc. Int. Conf. Learn. Represent.*, Addis Ababa, Ethiopia, 2020, pp. 1–18.

[2] P. Gordaliza, E. Del Barrio, G. Fabrice, and J.-M. Loubes, "Obtaining fairness using optimal transport theory," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2357–2365.

[3] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2015, pp. 259–268.

[4] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3315–3323.

[5] S. Barocas, M. Hardt, and A. Narayanan, "Fairness in machine learning," in *Proc. NIPS Tutorial*, 2017, pp. 1–181.

[6] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil, "Empirical risk minimization under fairness constraints," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 2791–2801.

[7] R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa, "Wasserstein fair classification," 2019, *arXiv:1907.12059*. [Online]. Available: http://arxiv.org/abs/1907.12059

[8] L. Risser, Q. Vincenot, N. Couellan, and J.-M. Loubes, "Using Wasserstein-2 regularization to ensure fair decisions with neural-network classifiers," 2019, *arXiv:1908.05783*. [Online]. Available: http://arxiv.org/abs/1908.05783

[9] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proc. 26th Int. Conf. World Wide Web*, Apr. 2017, pp. 1171–1180.

[10] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 325–333.

[11] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3384–3393.

[12] H. Zhao, A. Coston, T. Adel, and G. J. Gordon, "Conditional learning of fair representations," 2019, *arXiv:1910.07162*. [Online]. Available: http://arxiv.org/abs/1910.07162

[13] M. Sugiyama, M. Krauledat, and K.-R. Müller, "Covariate shift adaptation by importance weighted cross validation," *J. Mach. Learn. Res.*, vol. 8, pp. 985–1005, May 2007.

[14] W. Lee, J. Lee, and S. Park, "Instance weighting domain adaptation using distance kernel," *Ind. Eng. Manage. Syst.*, vol. 17, no. 2, pp. 334–340, Jun. 2018.

[15] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2200–2207.

[16] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, May 2015.

[17] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7167–7176.

[18] S. Park, W. Lee, and J. Lee, "Learning of indiscriminate distributions of document embeddings for domain adaptation," *Intell. Data Anal.*, vol. 23, no. 4, pp. 779–797, Sep. 2019.

[19] Y. Zhang, N. Wang, S. Cai, and L. Song, "Unsupervised domain adaptation by mapped correlation alignment," *IEEE Access*, vol. 6, pp. 44698–44706, 2018.

[20] F. Sun, H. Wu, Z. Luo, W. Gu, Y. Yan, and Q. Du, "Informative feature selection for domain adaptation," *IEEE Access*, vol. 7, pp. 142551–142563, 2019.

[21] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 137–144.

[22] M. Zhang, D. Wang, W. Lu, J. Yang, Z. Li, and B. Liang, "A deep transfer model with wasserstein distance guided multi-adversarial networks for bearing fault diagnosis under different working conditions," *IEEE Access*, vol. 7, pp. 65303–65318, 2019.

[23] A. Coston, K. N. Ramamurthy, D. Wei, K. R. Varshney, S. Speakman, Z. Mustahsan, and S. Chakraborty, "Fair transfer learning with missing protected attributes," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Jan. 2019, pp. 91–98.

[24] C. Schumann, X. Wang, A. Beutel, J. Chen, H. Qian, and E. H. Chi, "Transfer of machine learning fairness across domains," 2019, *arXiv:1906.09688*. [Online]. Available: http://arxiv.org/abs/1906.09688

[25] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, Mar. 2012.

[26] L. V. Kantorovich, "On the translocation of masses," *J. Math. Sci.*, vol. 133, no. 4, pp. 1381–1382, Mar. 2006.

[27] C. Villani, *Optimal Transport: Old and New*, vol. 338. Berlin, Germany: Springer-Verlag, 2008.

[28] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*. [Online]. Available: http://arxiv.org/abs/1701.07875

[29] C.-Y. Lee, T. Batra, M. H. Baig, and D. Ulbricht, "Sliced wasserstein discrepancy for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10285–10295.

[30] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[31] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," 2017, *arXiv:1701.04862*. [Online]. Available: http://arxiv.org/abs/1701.04862

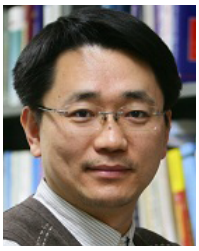
[32] J. Rabin, G. Peyré, J. Delon, and M. Bernot, "Wasserstein barycenter and its application to texture mixing," in *Proc. Int. Conf. Scale Space Variational Methods Comput. Vis.* Berlin, Germany: Springer, 2011, pp. 435–446.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[34] L. Ge, J. Gao, H. Ngo, K. Li, and A. Zhang, "On handling negative transfer and imbalanced distributions in multiple source transfer learning," *Stat. Anal. Data Mining, ASA Data Sci. J.*, vol. 7, no. 4, pp. 254–271, Aug. 2014.

[35] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

TAEHO YOON received the B.S. degree in mathematics from the Pohang University of Science and Technology, in 2018. He is currently pursuing the Ph.D. degree with the Department of Mathematical Sciences, Seoul National University, Seoul, South Korea. His research interests include mathematics for machine learning, information geometry, and optimization algorithms.



JAEWOOK LEE (Member, IEEE) received the B.S. degree in mathematics from Seoul National University, Seoul, South Korea, in 1993, and the Ph.D. degree in applied mathematics from Cornell University, in 1999. He is currently a Professor with the Department of Industrial Engineering, Seoul National University. His research interests include machine learning, neural networks, and global optimization and their applications to data mining and financial engineering.

WOOJIN LEE received the B.S. degree in industrial engineering from Yonsei University, in 2015. He is currently pursuing the Ph.D. degree with the Department of Industrial Engineering, Seoul National University, Seoul, South Korea. His research interests include domain adaptation, adversarial attacks, and fair machine learning.