Transfer of Machine Learning Fairness across Domains

Candice Schumann
University of Maryland
schumann@cs.umd.edu

Xuezhi Wang
Google
xuezhiw@google.com

Alex Beutel
Google
alexbeutel@google.com

Jilin Chen
Google
jilinc@google.com

Hai Qian
Google
hqian@google.com

Ed H. Chi
Google
edchi@google.com

Abstract

If our models are used in new or unexpected cases, do we know if they will make fair predictions? Previously, researchers developed ways to debias a model for a single problem domain. However, this is often not how models are trained and used in practice. For example, labels and demographics (sensitive attributes) are often hard to observe, resulting in auxiliary or synthetic data to be used for training, and proxies of the sensitive attribute to be used for evaluation of fairness. A model trained for one setting may be picked up and used in many others, particularly as is common with pre-training and cloud APIs. Despite the pervasiveness of these complexities, remarkably little work in the fairness literature has theoretically examined these issues.

We frame all of these settings as domain adaptation problems: how can we use what we have learned in a source domain to debias in a new target domain, without directly debiasing on the target domain as if it is a completely new problem? We offer new theoretical guarantees of improving fairness across domains, and offer a modeling approach to transfer to data-sparse target domains. We give empirical results validating the theory and showing that these modeling approaches can improve fairness metrics with less data.

1 Introduction

Much of machine learning research, and especially machine learning fairness, focuses on optimizing a model for a single use case [1, 4]. However, the reality of machine learning applications is far more chaotic. It is common for models to be used on multiple tasks, frequently different in a myriad of ways from the dataset that they were trained on, often coming at significant cost [24]. This is especially concerning for machine learning fairness – we want our models to obey strict fairness properties, but we may have far less data on how the models will actually be used. How do we understand our fairness metrics in these more complex environments?

In traditional machine learning, domain adaptation techniques are used when the distribution of training and validation data does not match the target distribution that the model will ultimately be tested against. Therefore, in this paper we ask: if the model is trained to be "fair" on one dataset, will it be "fair" over a different distribution of data? Instead of starting again with this new dataset,

can we use the knowledge gained during the original debiasing to more effectively debias in the new space?

It turns out that this framing covers many important cases for machine learning fairness. We will use, as a running example, the task of income prediction, where some decisions will be made based on the person's predicted income and we want the model to perform "fairly" over a sensitive attribute such as gender. We primarily follow the *equality of opportunity* [15] perspective where we are concerned with one group (broken down by gender or race) having worse accuracy than another. In this setting, there are a myriad of fairness issues that arise that we find domain adaptation can shed light on:

Lacking sensitive features for training: There may be few examples where we know the sensitive attribute. In these cases, a proxy of the sensitive attribute have been used [14], or researchers need very sample-efficient techniques [1, 4]. For distant proxies, researchers have asked how well fairness transfers across attributes [17]. Here the sensitive attribute differs in the source and target domains.

Data is not representative of application: Dataset augmentation, models offered as an API, or models used in multiple unanticipated settings, are all increasingly common design patterns. Even for machine learning fairness, researchers often believe limited training data is a primary source of fairness issues [7] and will employ dataset augmentation techniques to try to improve fairness [9]. How can we best make use of auxiliary data during training and evaluation when it differs in distribution from the real application?

Multiple tasks: In some cases having accurate labels for model training is difficult and instead proxy tasks with more labeled data are used to train the model, e.g., using pre-trained image or text models or using income brackets as a proxy for defaulting on a loan. Again we ask: when does satisfying a fairness property on the original task help satisfy that same property on the new task?

Each of these cases are common throughout machine learning but present challenges for fairness. In this work, we explore mapping domain adaptation principles to machine learning fairness. In particular, we offer the following contributions:

1. **Theoretical Bounds:** We provide theoretical bounds on transferring equality of opportunity and equality of odds metrics across domains. Perhaps more importantly, we discuss insights gained from these bounds.

2. **Modeling for Fairness Transfer:** We offer a general, theoretically-backed modeling objective that enables transferring fairness across domains.

3. **Empirical validation:** We demonstrate when transferring machine learning fairness works successfully, and when it does not, through both synthetic and realistic experiments.

2   Related Work

This work lies at the intersection of traditional domain adaptation and recent work on ML fairness.

Domain Adaptation   Both Pan et al. [23], and Weiss et al. [26] provide a survey on current work in transfer learning. One case of transfer learning is domain adaptation, where the task remains the same, but the distribution of features that the model is trained on (the source domain) does not match the distribution that the model is tested against (the target domain). Ben-David et al. [2] provide theoretical analysis of domain adaptation. Ben-David et al. [3] extend this analysis to provide a theoretical understanding of how much source and target data should be used to successfully transfer knowledge. Mansour et al. [22] provide theoretical bounds on domain adaptation using Rademacher Complexity analysis. In later research, Ganin et al. [11] build on this theory to use an adversarial training procedure over latent representations to improve domain adaptation.

Fairness in Machine Learning   A large thread of recent research has studied how to optimize for fairness metrics during model training. Li et al. [18] empirically show that adversarial learning helps preserve privacy over sensitive attributes. Beutel et al. [4] focus on using adversarial learning to optimize different fairness metrics, and Madras et al. [21] provides a theoretical framework for understanding how adversarial learning optimizes these fairness goals. Zhang et al. [28] use adversarial training over logits rather than hidden representations. Other work has focused on constraint-based optimization of fairness objectives [12, 1]. Tsipras et al. [25] however, provide a theoretical bound on the accuracy of adversarial robust models. They show that even with infinite data there will still be a trade-off of accuracy for robustness.

Domain Adaptation & Fairness Despite the prevalence of using one model across multiple domains, in practice little work has studied domain adaptation and transfer learning of fairness metrics. Kallus and Zhou [16] use covariate shift correction when computing fairness metrics to address bias in label collection. More related, Madras et al. [21] show empirically that their method allows for fair transfer. The transfer learning here corresponds to preserving fairness for a single sensitive attribute but over different tasks. However, Lan and Huan [17] found empirically that fairness does not transfer well to a new domain. They found that as accuracy increased in the transfer process, fairness decreases in the new domain. It is concerning that these papers show opposing effects. Both of these papers offer empirical results on the UCI adult dataset, but neither provide a theoretical understanding of how and when fairness in one domain transfers to another.

3 Problem Formulation

We begin with some notation to make precise the problem formulation. Building on our running example we have two domains: a source domain $Z \sim \mathcal{D}_S$, which is a feature distribution influenced by sensitive attribute $A_S \in \mathcal{A}_S$ (e.g., $\mathrm{Pr}_{Z \sim \mathcal{D}_S}[Z|A_S = \textit{male}] \neq \mathrm{Pr}_{Z \sim \mathcal{D}_S}[Z|A_S = \textit{female}]$), as well as a target domain $\mathcal{D}_T$ influenced by sensitive attribute $A_T \in \mathcal{A}_T$ (e.g., $\mathrm{Pr}_{Z \sim \mathcal{D}_T}[Z|A_T = \textit{black}] \neq \mathrm{Pr}_{Z \sim \mathcal{D}_T}[Z|A_T = \textit{white}]$). In order for this to be a domain adaptation problem, we assume $\mathrm{Pr}_{Z \sim \mathcal{D}_S}[Z|A_S] \neq \mathrm{Pr}_{Z \sim \mathcal{D}_T}[Z|A_T]$. Note, this can be true even if $\mathcal{D}_S = \mathcal{D}_T$ but the distributions conditioned on A_S and A_T differ. We focus on binary classification tasks with label $Y \in \mathcal{Y}$, e.g. income classification is shared over both domains. For this task we can create a classifier by finding a hypothesis $g : \mathcal{D} \to \mathcal{Y}$ from a hypothesis space $\mathcal{H}$.

Let us assume that we can learn a "fair" classifier g for the source domain and task. If we use a small amount of data from the target domain, will the fairness from the source sensitive attribute A_S transfer to the target domain and sensitive attribute A_T? We can define the notion of a "fairness" distance – how far away the classifier is from perfectly fair – in a given domain S as Δ_{Fair_S}. Within this formulation we consider two definitions of fairness.

The first distance is *equality of opportunity* [15]. A classifier is said to be fair under equality of opportunity if the false positive rates (FPR) over sensitive attributes are equal. In other words if we have a binary sensitive attribute A, then equality of opportunity requires that $\mathrm{Pr}(\hat{Y} = 1|A = 0, Y = 0) = \mathrm{Pr}(\hat{Y} = 1|A = 1, Y = 0)$, where $\hat{Y}$ gives the outcome of classifier g. Thus, how far away a classifier g is from equal opportunity (or the fairness distance of equal opportunity) can be defined as

$$\Delta_{EOp_S}(g) \triangleq \left| \mathbb{E}_{Z_0^0 \sim \mathcal{D}_{S_0^0}}[g(Z_0^0)] - \mathbb{E}_{Z_1^0 \sim \mathcal{D}_{S_1^0}}[g(Z_1^0)] \right|,$$

where $\mathcal{D}_{S_\alpha^l} = P_{Z \sim \mathcal{D}_S}[Z|A = \alpha, Y = l]$. In our running example $\Delta_{EOp_S}(g)$, where A_S is gender, is the difference between the likelihood that a low-income man is predicted to be high-income and the likelihood that a low-income woman is predicted to be high-income. A symmetric definition and set of analysis can be made for false negative rate (FNR).

The second definition of fairness which we consider is *equalized odds* [15]. A classifier is said to be fair under equalized odds if both the FPR *and* FNR over the sensitive attribute are equal: Similar to equal opportunity, we define the fairness distance of equalized odds as:

$$\Delta_{EO_S}(g) \triangleq \left| \mathbb{E}_{Z_0^0 \sim \mathcal{D}_{S_0^0}}[g(Z_0^0)] - \mathbb{E}_{Z_1^0 \sim \mathcal{D}_{S_1^0}}[g(Z_1^0)] \right| + \left| \mathbb{E}_{Z_0^1 \sim \mathcal{D}_{S_0^1}}[1 - g(Z_0^1)] - \mathbb{E}_{Z_1^1 \sim \mathcal{D}_{S_1^1}}[1 - g(Z_1^1)] \right|.$$

Again using our running example, the distance of equalized odds in the source domain is given by the difference of expected FPRs between females and males (as above), plus the difference of expected FNRs (high-income predicted to be low-income) between females and males.

Given a classifier g that has a fairness guarantee in the source domain, the fairness distance in the target domain should be bounded by the fairness distance in the source domain:

$$\Delta_{Fair_T}(g) \leq \Delta_{Fair_S}(g) + \epsilon \tag{1}$$

The key question we hope to answer is: what is ϵ?

4 Bounds on Fairness in the Target Domain

To expand inequality (1) we need to start with some definitions. Given a hypothesis space $\mathcal{H}$ and a true labeling function $f(Z) : \mathcal{D} \to \mathcal{Y}$, we can define the error of a hypothesis $g \in \mathcal{H}$ as $\epsilon_S(g, f) = \mathbb{E}_{Z \sim \mathcal{D}_S}[|f(Z) - g(Z)|]$, the expectation of disagreement between the hypothesis g and the true label f. We can then define the ideal joint hypothesis that minimizes the combined error over both the source and target domains as $g^* = \arg\min_{g \in \mathcal{H}} \epsilon_S(g, f) + \epsilon_T(g, f)$.

Following Ben-David et al. [3] we define the $\mathcal{H}$-divergence between probability distributions as

$$d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') = 2 \sup_{g \in \mathcal{H}} |\Pr_{\mathcal{D}}[I(g)] - \Pr_{\mathcal{D}'}[I(g)]|, \tag{2}$$

where $I(g)$ is the set for which $g \in \mathcal{H}$ is the characteristic function ($Z \in I(g) \Leftrightarrow g(Z) = 1$). We can compute an approximation $\hat{d}_{\mathcal{H}}(\mathcal{D}, \mathcal{D}')$ by finding a hypothesis h that finds the largest difference between the samples from $\mathcal{D}$ and $\mathcal{D}'$ [2]. This divergence can be used to look at the differences in distributions, which is important when moving from a source domain to a target domain.

Additionally, we defined the symmetric difference hypothesis space $\mathcal{H}\Delta\mathcal{H}$ as the set of hypotheses

$$g \in \mathcal{H}\Delta\mathcal{H} \iff g(Z) = h(Z) \oplus h'(Z) \quad \text{for some } h, h' \in \mathcal{H}, \tag{3}$$

where $\oplus$ is the XOR function. The symmetric difference hypothesis space is used to find disagreements between a potential classifier g and a true labeling function f.

Theorem 1. *Let $\mathcal{H}$ be a hypothesis space of VC dimension d. If $\mathcal{U}_{S_0^0}$, $\mathcal{U}_{S_1^0}$, $\mathcal{U}_{T_0^1}$, $\mathcal{U}_{T_1^0}$ are samples of size m', each drawn from $\mathcal{D}_{S_0^0}$, $\mathcal{D}_{S_1^0}$, $\mathcal{D}_{T_0^0}$, and $\mathcal{D}_{T_1^0}$ respectively, then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ (over the choice of samples), for every $g \in \mathcal{H}$ (where $\mathcal{H}$ is a symmetric hypothesis space) the distance from equal opportunity in the target space is bounded by*

$$\Delta_{EOp_T}(g) \leq \Delta_{EOp_S}(g) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^0}, \mathcal{U}_{S_0^0}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^0}, \mathcal{U}_{S_1^0})$$

$$+ 8\sqrt{\frac{2d\log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda_0^0 + \lambda_1^0,$$

where $\lambda_\alpha^l = \epsilon_{S_\alpha^l}(g^*, f) + \epsilon_{T_\alpha^l}(g^*, f)$.

Using both the definition of $\mathcal{H}$-divergence and symmetric difference hypothesis space, Theorem 1 provides a VC-dimension bound on the equal opportunity distance in the target domain given the equal opportunity distance in the source domain. Due to space limitations, full proofs for all theorems can be found in Appendix B.

This theorem provides insights on when domain adaptation for fairness can be used. Firstly the $\hat{d}$ terms in the bound suggest that 1) the source and target distributions of negatively labeled items that have a sensitive attribute label of 0 should be close, and 2) the source and target distributions of the negatively labeled items that have a sensitive attribute label of 1 should be close. In Figure 1 the red quadrants should be close to the red quadrants while the orange quadrants should be close to the orange quadrants
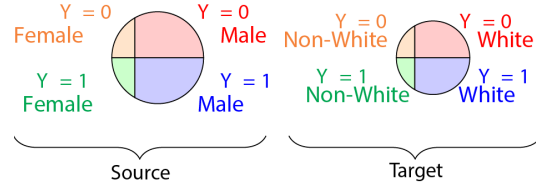

Figure 1: Both the source and target distributions can be split into four quadrants: 1) negative minority 2) negative majority 3) positive minority and 4) positive majority.

rants across domains. In traditional domain adaptation, ignoring fairness, the entire domains should be close (the entire circle), which means that if there are few minority data-points then the distance of the minority spaces will be ignored. The fairness bound instead puts equal emphasis on both the majority and minority.

Secondly, the λ terms become small when the hypothesis space contains a function g^* that has low error on both the source and target space on the two negative segments in each domain (the red and orange spaces in Figure 1). Since we are looking at equal opportunity, the function g^* only needs to have low error on the negative space for both the majority and minority. Therefore, we can use the trivial function $g^*(Z) = 0$ and the λ terms go to 0.

Lastly, Theorem 1 depends on the VC-dimension d. Since bounds with VC-dimensions explode with models like neural networks, we also provide bounds using Rademacher Complexity in Appendix A.

Equalized odds, while similar to equal opportunity, is a stricter fairness constraint. Theorem 2 provides a VC-dimension bound on the difference of equal odds in the target domain given the source domain.

Theorem 2. *Let $\mathcal{H}$ be a hypothesis space of VC dimension d. If $\mathcal{U}_{S_\alpha^l}$ are samples of size m', each drawn from $\mathcal{D}_{S_\alpha^l}$ for all $\alpha \in \mathcal{A} = \{0,1\}$ and $l \in \mathcal{Y} = \{0,1\}$, then for any $\delta \in (0,1)$, with probability at least $1 - \delta$ (over the choice of samples), for every $g \in \mathcal{H}$ (where $\mathcal{H}$ is a symmetric hypothesis space) the distance from equalized odds in the target space is bounded by*

$$\Delta_{EO_T}(g) \le \Delta_{EO_S}(g) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^0}, \mathcal{U}_{S_0^0}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^0}, \mathcal{U}_{S_1^0})$$

$$+ \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^1}, \mathcal{U}_{S_0^1}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^1}, \mathcal{U}_{S_1^1}) + 16\sqrt{\frac{2d\log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda_{EO},$$

where $\lambda_{EO} = \lambda_0^0 + \lambda_1^0 + \lambda_0^1 + \lambda_1^1$, and $\lambda_\alpha^l = \epsilon_{S_\alpha^l}(g^*, f) + \epsilon_{T_\alpha^l}(g^*, f)$.

The $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}$ terms suggest, that in order for equalized odds to transfer successfully then, 1) the source and target distributions of negatively labeled items on *both* sensitive attribute labels 0 and 1 should be close, 2) the source and target distributions of the positively labeled items on *both* sensitive attribute labels 0 and 1 should be close. In other words, all four quadrants of the source should individually be close to the respective four quadrants of the target in Figure 1.

Additionally, the λ term shows that there should be a hypothesis that performs well over *all* of these subspaces. This implication is intuitive given that equalized odds, by definition, wants a classifier to perform well in both the negative and positive space across both groups.

5 Modeling to Transfer Fairness

With this theoretical understanding, how should we change our training? As motivated previously, we consider the case where we have a small amount of labelled data (both labels $\mathcal{Y}$ and sensitive attributes $\mathcal{A}$) in the target domain and a large amount of labelled data in the source domain.

As shown in the previous section, equality of opportunity will transfer *if* the distance between the respective distributions of source and target are close together as visually portrayed in Figure 1. Ganin et al. [11] proved that traditional domain adaptation can be framed as minimizing the distance between source and target with adversarial training. [20, 10, 4, 18] similarly have applied adversarial training to achieve fairness goals, and Madras et al. [21] proved that equality of odds can be optimized with adversarial training similar to domain adaptation.
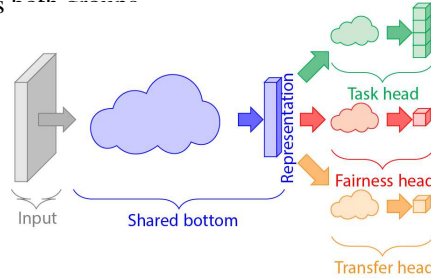

Figure 2: At a high level, our general framework combines a primary training objective, a fairness objective, and a transfer objective to improve fairness goals in a target domain. Table 1 provides mathematical details for different configurations.

We build on this intuition to design a learning objective for transferring equality of opportunity to a target domain. Adversarial training conceptually enables minimizing a $\hat{d}$ term from Theorem 1; and Δ_{Fair_S} can be optimized using [4, 21] or one of the other myriad of traditional fairness learning objectives. As such, we begin with the following loss:

$$\min \left[\sum_{Z \sim (\mathcal{D}_S \cup \mathcal{D}_T)} L_Y(g(h(Z)), f(Z)) + \sum_{(A, Z^0) \sim \mathcal{D}_{S^0}} \lambda_{Fair} L_A(a(h(Z^0)), A) \right.$$

$$\left. + \sum_{(d, Z_0^0) \sim \left(\mathcal{D}_{S_0^0} \cup \mathcal{D}_{T_0^0}\right)} \lambda_{DA} L_d(d(h(Z_0^0)), d) + \sum_{(d, Z_1^0) \sim \left(\mathcal{D}_{S_1^0} \cup \mathcal{D}_{T_1^0}\right)} \lambda_{DA} L_d(d(h(Z_1^0)), d) \right], \quad (4)$$

where $L_Y(g(h(Z)), f(Z))$ is the loss function training $g(h(Z))$ over hidden representation $h(Z)$ to predict the task label $f(Z)$. To optimize Δ_{Fair_S}, $a(h(Z^0))$ tries to predict the sensitive attribute A from the source and $L_A(a(h(Z^0)), A)$ provides an adversarial loss that includes a negated gradient on h following [4]. For transfer, we minimize $\hat{d}$ terms by including another adversarial loss

Loss Term	Theorem 1	Adversarial (Eq. 4)	Regularization (Eq. 5)
Fairness head	$\Delta_{EOp_S}(g)$	$\lambda_{Fair}L_A\left(a(h(Z^0)),A\right)$	$\lambda_{Fair}L_{MMD}\left(a(h(Z^0)),A\right)$
Transfer head	$\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^0},\mathcal{U}_{S_0^0})$	$\lambda_{DA}L_d\left(d(h(Z_0^0)),d\right)$	$\lambda_{DA}L_{MMD}\left(d(h(Z^0)),d\right)$
	$\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^0},\mathcal{U}_{S_1^0})$	$\lambda_{DA}L_d\left(d(h(Z_1^0)),d\right)$	

Table 1: Relationship between terms in Theorem 1 and Loss functions

$L_d\left(d(h(Z_l^\alpha)),d\right)$, where $d(h(Z_l^\alpha))$ tries to predict whether a sample comes from the source or target domain. Each of these loss components maps to terms in Theorem 1 as laid out in Table 1.

Recently, Zhang et al. [28] used adversarial training on a one dimensional representation of the data (effectively the model's prediction). From this perspective, we can use a wide variety of losses over predictions to replace adversarial losses, such as [27, 5] minimizing the correlation between group and the one dimensional representation of the data. Like previous work, we find that these approaches to be more stable and still effective in comparison to adversarial training, despite not being provably optimal. In our experiments we use a MMD loss [13, 19, 6] over predictions:

$$\min\left[\sum_{Z\in\mathcal{D}_S\cup\mathcal{D}_T}L_Y(f(Z),g(Z)) + \sum_{(A,Z^0)\sim\mathcal{D}_{S^0}}\lambda_{Fair}L_{MMD}\left(a(h(Z^0)),A\right)\right.$$

$$\left.+ \sum_{(d,Z^0)\sim\left(\mathcal{D}_{S^0}\cup\mathcal{D}_{T^0}\right)}\lambda_{DA}L_{MMD}\left(d(h(Z^0)),d\right)\right], \tag{5}$$

where $\lambda_{Fair}L_{MMD}\left(a(h(Z^0)),A\right)$ is the MMD regularization over the sensitive attributes in the source domain, $\lambda_{DA}L_{MMD}\left(d(h(Z^0)),d\right)$ is the MMD regularization over source/target membership. Again Table 1 maps the terms in Eq. 5 to those in Theorem 1.

Care must be taken when performing domain adaptation with regards to fairness. Either multiple transfer heads should be included in the loss for all necessary quadrants (See Figure 1 and Eq. 4), or balanced data – equally representing all necessary quadrants – should be used as in [21] and Eq. 5. Experiments in this paper use the MMD regularization as in Eq. 5 and balanced data is used for both the fairness head as well as the transfer heads.

6 Experiments

To better understand the theoretical results presented above, we now present both synthetic and realistic experiments exploring tightness of our theoretical bound as well as the ability to improve the transfer of fairness across domains during model training.

6.1 Synthetic Examples

(a) Source (b) Target -1 (c) Target 0 (d) Target 1 (e) Target Fairness

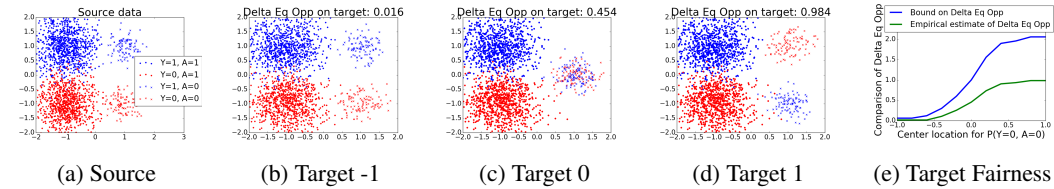Figure 3: Synthetic examples showing how distribution difference of $P(Z|Y,A=0)$ in the target domain affects theoretical and empirical equality of opportunity (best viewed in color). In the title of each plot we give the equal opportunity distance $\Delta_{EOp_T}(g)$ in the target domain.

We show how well the theoretical bounds align with actual transfer of fairness. A synthetic dataset is used to examine how the distribution distance terms $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_{A=0}^{Y=0}},\mathcal{U}_{S_{A=0}^{Y=0}})$ and $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_{A=1}^{Y=0}},\mathcal{U}_{S_{A=1}^{Y=0}})$ in Eq. (1) affect the fairness distance of equal opportunity $\Delta_{EOp_T}(g)$.

In this synthetic example, we generate data $Z\in\mathbb{R}^2$ using Gaussian distributions. As we can see in Figure 3a, the source domain consists of four Gaussians, with $Y=1$ largely lying above $Y=0$ and $A=1$ lying to the left of $A=0$; $A=1$ is the majority of the data ($\sigma=0.5$ with 900 samples). For $A=0$, the data is generated using $\sigma=0.3$ with 100 samples. The target domain, like the source
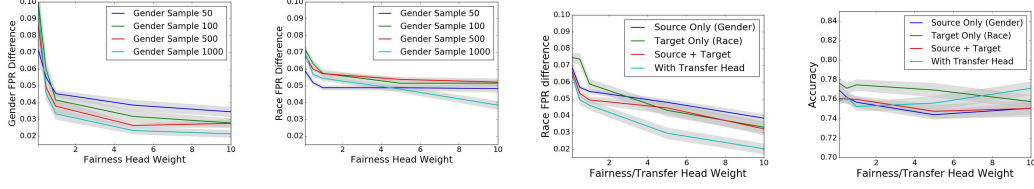
(a) Effect of fairness head: Improving $\Delta_{EOp_{\text{gender}}}$ with varying number of gender-balanced samples.

(b) Some natural transfer occurring without explicit transfer: $\Delta_{EOp_{\text{race}}}$ is improved with gender data.

(c) Effect of transfer head: better transfer from gender (1000 samples) to race (50 samples).

(d) Accuracy graph for transferring from gender (1000 samples) to race (50 samples).

Figure 4: Effect of fairness/transfer head on the UCI data. The shaded areas show the standard error of the mean across trials. Note the head weight (x-axis) starts from 0.1.

domain, consists of majority data with $A = 1$ and the data from $A = 1$ is generated from the same distribution in both domains: $\mathcal{U}_{T_{A=1}^{Y=0}} \sim \mathcal{N}([-1,-1],\sigma)$ and $\mathcal{U}_{T_{A=1}^{Y=1}} \sim \mathcal{N}([-1,1],\sigma)$. However, in order to understand the transfer of fairness, we shift the distributions of $\mathcal{U}_{T_{A=0}^{Y=0}} \sim \mathcal{N}([1,c],\sigma)$ and $\mathcal{U}_{T_{A=0}^{Y=1}} \sim \mathcal{N}([1,-c],\sigma)$ in the target domain ($c = -1,0,1$ for 3b, 3c and 3d, respectively). By varying the overlap between these distributions, and their alignment with the source data, we are able to understand the relationship between the $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}$ terms above and the fairness distance of equal opportunity $\Delta_{EOp_T}(g)$. For each setting, we train linear classifiers on the source domain and examine the performance in the target domain.

Qualitative Analysis We see in Fig. 3b that when the distribution $P(Z|Y = 0, A = 0)$ across domains is close, thus a smaller $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^0}, \mathcal{U}_{S_0^0})$, there is better transfer of fairness the source to the target domain, seen in the smaller $\Delta_{EOp_T}(g)$. As the distribution distance gets larger, the $\Delta_{EOp_T}(g)$ also increases. Consider the worst case of a sign flip for the minority $A = 0$, as shown in Fig. 3d: the FPR for the majority $A = 1$ is close to 0%, while the FPR for the minority $A = 0$ is close to 100%.

Quantitative Analysis In Figure 3e, we compare the derived bound of $\Delta_{EOp_T}(g)$ (Eq. 1) with its empirical estimate as we vary c[1]. As shown in Figure 3e, the theoretical bound on the equal opportunity distance is close to the observed equal opportunity distance when the distance between the negative minority space across domains, $\hat{d}(\mathcal{U}_{T_0^0}, \mathcal{U}_{S_0^0})$, is small. This suggests, minimizing the domain distance terms in Eq. 1 could lead to a better equal opportunity transfer.

6.2 Real Data

We now explore how and when our proposed modeling approach in Section 5 facilitates the transfer of fairness from the source to the target domain on two real-world datasets. Note, we use these datasets exclusively for understanding our theory and model, and *not* as a comment on when or if the proposed tasks and their application are appropriate, as in [1].

Dataset 1: The UCI Adult[2] dataset contains census information of over 40,000 adults from the 1994 Census, with the task of determining income brackets of $> \$50,000$ or $\leq \$50,000$. We focus on two sensitive attributes: binary valued gender, and race, converted to binary values ['white', 'non-white'] as done by Madras et al. [21].

Dataset 2: As in [1] we use ProPublica's COMPAS recidivism data[3] to try to predict recidivism for over 10,000 defendants based on age, gender, demographics, prior crime count, etc. We again focus on two sensitive attributes: gender and race (binarized to ['white', 'non-white']).

Experiment Setup For both datasets, cross-validation is used to choose the hyper-parameters. Comparable baseline accuracy (around 84% for Dataset 1 and 80% for Dataset 2, see appendix D for more details) is achieved with 64 embedding dimension for categorical features, single hidden layer with 256 shared hidden units, 512 batch size, 0.1 learning rate with Adagrad optimizer, and

---

[1]As in [2], $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^0}, \mathcal{U}_{S_0^0})$ is estimated by a linear classifier trained on samples $\mathcal{U}_{T_0^0}, \mathcal{U}_{S_0^0}$. The plot omits the VC term for simplicity, which is relatively small when sample size m' is large and VC-dimension d is low.

[2]https://archive.ics.uci.edu/ml/datasets/adult

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

$10,000$ epochs for training. We perform 30 runs for each set of experiments and average over the results.

Sparsity Issues and Natural Transfer   We examine the effectiveness of just the fairness heads in the proposed model. The amount of gender-balanced data created for the fairness head is varied to observe how applying the fairness head affects the FPR difference.

We examine how this procedure effects the FPR difference across genders (i.e., the FPR difference between "Female" and "Male" examples). Figure 4a shows that the fairness head works as expected: with sufficient data and a large enough weight, the fairness head is able to improve the FPR gap across genders. Further, we find that with very few examples on which to apply the fairness head, the gender FPR gap does not close. This aligns with previous results found in [4, 21, 5].

Second, we examine how running the fairness head on gender affects the FPR gap across race. As shown in Figure 4b, there is a natural transfer of equal opportunity from gender to race – applying a fairness loss with respect to gender also improves the fairness of the model with respect to race. This highlights that sometimes there is a natural transfer of equal opportunity, presenting general value in improving the FPR gap with respect to gender, and no explicit transfer optimization is needed. (Similar to the transfer questions posed previously by Madras et al. [21] and Gupta et al. [14]).

Effectiveness of Transfer Head   We now explore how adding the transfer head can further improve equality of opportunity in the target domain. We compare four different model arrangements: (1) **Source Only**: We only add a fairness head for the source domain; (2) **Target Only**: We only add a fairness head for the target domain; (3) **Source+Target**: We add two fairness heads, one for source and for target; (4) **Transfer**: We include three heads – both source and target fairness heads as well as the transfer head for equality of opportunity.

Experiment setting: As in typical transfer learning setting, we will focus on the case where we observe a large number of samples in the source domain (e.g., 1000 for each race "white" and "non-white"), but a smaller sample size in the target domain (e.g., 100 for each gender "male" and "female"), and the same for gender to race. We explore equality of opportunity with respect to FPR in the target domain, as we vary the weight on the fairness and transfer heads.

Results: Figure 4c shows that including the transfer head results in a better equal opportunity transfer, compared to the same setting without transfer (Figure 4b). Table 2 summarizes the full results on both datasets. We can see that including both the fairness heads and the transfer head consistently gives the best improvement in equal opportunity (FPR difference) in almost all cases.

Effect of Target Sample Size   Last, we consider how the amount of data from the target domain affects our ability to improve equal opportunity there, as sample efficiency is a core challenge.

Experiment setting: We follow a similar experimental procedure as before with two modifications. First, we vary the number of samples we observe for each sensitive group in the target domain to be in $\{50, 100, 500, 1000\}$. We examine the efficacy of the four approaches depending on the amount of data available for debiasing in the target domain. Second, this analysis is performed for both transferring from race (source) to gender (target), as well as from gender (source) to race (target).

Results: Table 2 summarizes the results. Applying the fairness and transfer heads to the large amount of source data closes the FPR gap in the target domain. Increasing the amount of data in the target domain significantly helps the performance of the "Target Only" and the "Source+Target" models. This is intuitive since directly debiasing in the target domain is feasible with sufficient data. With sufficient data, the results converge to be approximately equivalent to the transfer model.

These experiments show that the transfer model is effective in decreasing the FPR gap in the target domain and is more sample efficient than previous methods.

7   Conclusion

In this paper we provide the first theoretical examination of transfer of machine learning fairness across domains. We adopt a general formulation of domain adaptation for fairness that covers a wide variety of fairness challenges, from proxies of sensitive attributes, to applying models in unanticipated settings. Within this general formulation, we have provided theoretical bounds on the transfer of fairness for equal opportunity and equalized odds using both VC-dimension and Rademacher Complexity. Based on this theory, we developed a new modeling approach to transfer fairness to

			Smallest FPR difference achieved on Target (FPR-diff $\pm$ std. dev)			
	Source to Target	#Target Samples	Source only	Target only	Source + Target	With Transfer Head
Dataset 1	Gender to Race	50	0.038 ± 0.013	0.033 ± 0.019	0.032 ± 0.020	$\mathbf{0.020 \pm 0.016}$
		100	$\mathbf{0.038 \pm 0.013}$	$\mathbf{0.038 \pm 0.021}$	0.044 ± 0.024	0.040 ± 0.024
		500	0.038 ± 0.013	0.053 ± 0.010	0.043 ± 0.017	$\mathbf{0.025 \pm 0.018}$
		1000	0.038 ± 0.013	$\mathbf{0.027 \pm 0.018}$	0.027 ± 0.019	0.031 ± 0.021
	Race to Gender	50	0.061 ± 0.054	0.035 ± 0.015	0.020 ± 0.026	$\mathbf{0.008 \pm 0.009}$
		100	0.061 ± 0.054	0.028 ± 0.014	0.021 ± 0.015	$\mathbf{0.009 \pm 0.011}$
		500	0.061 ± 0.054	0.028 ± 0.013	0.019 ± 0.013	$\mathbf{0.014 \pm 0.011}$
		1000	0.061 ± 0.054	0.021 ± 0.012	$\mathbf{0.015 \pm 0.014}$	0.020 ± 0.014
Dataset 2	Gender to Race	50	0.027 ± 0.008	0.041 ± 0.006	0.009 ± 0.004	$\mathbf{0.001 \pm 0.001}$
		100	0.027 ± 0.008	0.036 ± 0.007	0.005 ± 0.005	$\mathbf{0.003 \pm 0.001}$
		500	0.027 ± 0.008	0.038 ± 0.008	0.003 ± 0.002	$\mathbf{0.001 \pm 0.001}$
		1000	0.027 ± 0.008	0.021 ± 0.005	0.006 ± 0.005	$\mathbf{0.002 \pm 0.001}$
	Race to Gender	50	0.040 ± 0.004	0.070 ± 0.005	0.035 ± 0.004	$\mathbf{0.019 \pm 0.002}$
		100	0.040 ± 0.004	0.055 ± 0.007	0.034 ± 0.003	$\mathbf{0.017 \pm 0.002}$
		500	0.040 ± 0.004	0.042 ± 0.008	0.027 ± 0.004	$\mathbf{0.019 \pm 0.002}$
		1000	0.040 ± 0.004	0.034 ± 0.011	0.028 ± 0.004	$\mathbf{0.018 \pm 0.002}$

Table 2: Comparison between the proposed model and the baselines. The numbers in bold indicate the smallest FPR difference achieved in the target domain w.r.t. varying number of target samples.

a given target domain. In experiments we validate our theoretical results and demonstrate that our modeling approach is more sample efficient in improving fairness metrics in a target domain.

References

[1] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. M. Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 60–69, 2018.

[2] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144, 2007.

[3] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[4] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. *Proceedings of the Conference on Fairness, Accountability and Transparency*, 2017.

[5] A. Beutel, J. Chen, T. Doshi, H. Qian, A. Woodruff, C. Luu, P. Kreitmann, J. Bischof, and E. H. Chi. Putting fairness principles into practice: Challenges, metrics, and improvements. *Artificial Intelligence, Ethics, and Society*, 2019.

[6] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 343–351, 2016.

[7] I. Chen, F. D. Johansson, and D. Sontag. Why is my classifier discriminatory? *arXiv preprint arXiv:1805.12002*, 2018.

[8] K. Crammer, M. Kearns, and J. Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9(Aug):1757–1774, 2008.

[9] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. Measuring and mitigating unintended bias in text classification. In *available at: www. aies-conference. com/wp-content/papers/main/AIES_2018_paper_9. pdf (accessed 6 August 2018).[Google Scholar]*, 2018.

[10] H. Edwards and A. J. Storkey. Censoring representations with an adversary. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[11] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[12] G. Goh, A. Cotter, M. Gupta, and M. P. Friedlander. Satisfying real-world goals with dataset constraints. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2415–2423. Curran Associates, Inc., 2016.

[13] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. In *The Journal of Machine Learning Research*, 2012.

[14] M. R. Gupta, A. Cotter, M. M. Fard, and S. Wang. Proxy fairness. *CoRR*, abs/1806.11212, 2018. URL http://arxiv.org/abs/1806.11212.

[15] M. Hardt, E. Price, N. Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[16] N. Kallus and A. Zhou. Residual unfairness in fair machine learning from prejudiced data. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 2444–2453, 2018.

[17] C. Lan and J. Huan. Discriminatory transfer. *CoRR*, 2017. URL http://arxiv.org/abs/1707.00780.

[18] Y. Li, T. Baldwin, and T. Cohn. Towards robust and privacy-preserving text representations. *arXiv preprint arXiv:1805.06093*, 2018.

[19] M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, 2015.

[20] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. S. Zemel. The variational fair autoencoder. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

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

[22] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *COLT*, 2009.

[23] S. J. Pan, Q. Yang, et al. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[24] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison. Hidden technical debt in machine learning systems. In *Advances in neural information processing systems*, pages 2503–2511, 2015.

[25] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. There is no free lunch in adversarial robustness (but there are unexpected benefits). *arXiv preprint arXiv:1805.12152*, 2018.

[26] K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big Data*, 2016.

[27] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, pages 962–970, 2017.

[28] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. *CoRR*, abs/1801.07593, 2018. URL http://arxiv.org/abs/1801.07593.

A Rademacher Complexity

We provide additional bounds dependent on Radmacher Complexity based on the following definition of data-driven empirical Rademacher Complexity

Definition 1. Given a hypothesis space $\mathcal{H}$, a sample $S \in \mathcal{X}^m$, the empirical Rademacher Complexity of $\mathcal{H}$ is defined as

$$\hat{\mathfrak{R}}_S(\mathcal{H}) = \frac{2}{m}\mathbb{E}_\sigma \left[\sup_{h\in\mathcal{H}} |\sum_{i=1}^m \sigma_i h(x_i)|\,\middle|\, S = (x_1,\ldots,x_m)\right].$$

The expectation is taken over $\sigma = (\sigma_1,\ldots,\sigma_m)$ where $\sigma_i \in \{-1,+1\}$ are uniform independent random variables. The Rademacher Complexity of a hypothesis space is defined as the expectation of $\hat{\mathfrak{R}}$ over all sample sets of size m

$$\mathfrak{R}_m(\mathcal{H}) = \mathbb{E}_S \left[\hat{\mathfrak{R}}_S(\mathcal{H})\,\middle|\, |S| = m\right]. \tag{6}$$

Rademacher Complexity measures the ability of a hypothesis space to fit random noise. The empirical Rademacher Complexity function allows us to estimate the Rademacher Complexity using a finite sample of data. Rademacher Complexity bounds can lead to tighter bounds than those of VC-dimension, especially when analyzing neural network models.

When transitioning to Rademacher Complexity we need to change the binary labels from $\{0,1\}$ to $\{-1,1\}$. This means that the error of a hypothesis g is defined as

$$\epsilon_{S_\alpha^l}(g,f) = \mathbb{E}_{z_\alpha^l \sim D_{S_\alpha^l}} \left[\frac{|g(z_\alpha^l) - f(z_\alpha^l)|}{2}\right].$$

Additionally, we need new definitions of the equal opportunity and equalized odds distances over the new binary group membership. The equal opportunity distance is defined as

$$\Delta_{EOp_S}(g) \triangleq \mathbb{E}_{Z_0^{-1} \sim D_{S_0^{-1}}} \left[\frac{1 + g(z_0^{-1})}{2}\right] - \mathbb{E}_{Z_1^{-1} \sim D_{S_1^{-1}}} \left[\frac{1 + g(z_1^{-1})}{2}\right],$$

while the equlized odds distance is defined as

$$\Delta_{EO_T}(g) \triangleq \left|\mathbb{E}_{Z_0^{-1} \sim D_{T_0^{-1}}} \left[\frac{1 + g(z_0^{-1})}{2}\right] - \mathbb{E}_{Z_1^{-1} \sim D_{T_1^{-1}}} \left[\frac{1 + g(z_1^{-1})}{2}\right]\right|$$
$$+ \left|\mathbb{E}_{Z_0^1 \sim D_{T_0^1}} \left[\frac{1 + g(z_0^1)}{2}\right] - \mathbb{E}_{Z_1^1 \sim D_{T_1^1}} \left[\frac{1 + g(z_1^1)}{2}\right]\right|.$$

Using these new definitions Theorem 3 provides a Rademacher Complexity bound of the equal opportunity distance in the target space. This closely resembles the VC-dimension bound in Theorem 1.

Theorem 3. *Let $\mathcal{H}$ be a hypothesis space. If $\mathcal{U}_{S_0^{-1}}$, $\mathcal{U}_{S_1^{-1}}$, $\mathcal{U}_{T_0^{-1}}$, $\mathcal{U}_{T_1^{-1}}$ are samples of size m', each drawn from $\mathcal{D}_{S_0^{-1}}$, $\mathcal{D}_{S_1^{-1}}$, $\mathcal{D}_{T_0^{-1}}$, and $\mathcal{D}_{T_1^{-1}}$ respectively, then for any $\delta \in (0,1)$, with probability at least $1 - \delta$ (over the choice of samples), for every $g \in \mathcal{H}$ (where $\mathcal{H}$ is a symmetric hypothesis space) the distance from equal opportunity in the target space is bounded by*

$$\Delta_{EOp_T}(g) \leq \Delta_{EOp_S}(g) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^{-1}}, \mathcal{U}_{S_0^{-1}}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^{-1}}, \mathcal{U}_{S_1^{-1}})$$
$$+ 2\left(\mathfrak{R}_{U_{T_0^{-1}}}(\mathcal{H}) + \mathfrak{R}_{U_{S_0^{-1}}}(\mathcal{H}) + \mathfrak{R}_{U_{T_1^{-1}}}(\mathcal{H}) + \mathfrak{R}_{U_{S_1^{-1}}}(\mathcal{H})\right)$$
$$+ 6\sqrt{\frac{\log\frac{2}{\delta}}{2m}} + \lambda_0^{-1} + \lambda_1^{-1},$$

where $\lambda_\alpha^l = \epsilon_{S_\alpha^l}(g^*, f) + \epsilon_{T_\alpha^l}(g^*, f)$.

The proof also follows a similar logic to the sketch given for Theorem 1 with the additional step of using a modification of Corollary 7 given by Mansour et al. [22].

Similarly, Theorem 4 provides a Rademacher Complexity bound of the equalized odds distance in the target space.

Theorem 4. *Let $\mathcal{H}$ be a hypothesis space. If $\mathcal{U}_{S_0^{-1}}$, $\mathcal{U}_{S_1^{-1}}$, $\mathcal{U}_{T_0^{-1}}$, $\mathcal{U}_{T_1^{-1}}$ $\mathcal{U}_{S_0^1}$, $\mathcal{U}_{S_1^1}$, $\mathcal{U}_{T_0^1}$, $\mathcal{U}_{T_1^1}$ are samples of size m', each drawn from $\mathcal{D}_{S_0^{-1}}$, $\mathcal{D}_{S_1^{-1}}$, $\mathcal{D}_{T_0^{-1}}$, $\mathcal{D}_{T_1^{-1}}$, $\mathcal{D}_{S_0^1}$, $\mathcal{D}_{S_1^1}$, $\mathcal{D}_{T_0^1}$, and $\mathcal{D}_{T_1^1}$ respectively, then for any $\delta \in (0,1)$, with probability at least $1 - \delta$ (over the choice of samples), for every $g \in \mathcal{H}$ (where $\mathcal{H}$ is a symmetric hypothesis space) the distance from equalized odds in the target space is bounded by*

$$\Delta_{EO_T}(g) \leq \Delta_{EO_S}(g) + \frac{1}{2}\left(\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_0^{-1}},\mathcal{U}_{T_0^{-1}}) + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_1^{-1}},\mathcal{U}_{T_1^{-1}})\right.$$
$$\left. +\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_0^1},\mathcal{U}_{T_0^1}) + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_1^1},\mathcal{U}_{T_1^1})\right)$$
$$+ 2\left(\hat{\mathfrak{R}}_{U_{S_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{S_1^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_1^{-1}}}(\mathcal{H})\right.$$
$$\left. + \hat{\mathfrak{R}}_{U_{S_0^1}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_0^1}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{S_1^1}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_1^1}}(\mathcal{H})\right)$$
$$+ 12\sqrt{\frac{\log\frac{2}{\delta}}{2m}} + \lambda_{EO},$$

where $\lambda_{EO} = \lambda_0^{-1} + \lambda_1^{-1} + \lambda_0^1 + \lambda_1^1$, and $\lambda_\alpha^l = \epsilon_{S_\alpha^l}(g^*,f) + \epsilon_{T_\alpha^l}(g^*,f)$.

Given either the Rademacher Complexity bounds or the VC-dimension bounds, the implications stay the same. In order for a successful transfer of fairness the two (or four) subspace domains should be close across the source and target domains. Additionally, there should be a hypothesis in the hypothesis space that performs well over all of the relevant subspaces.

B Proofs

Lemma 1. *(From Ben-David et al. [3]) For any hypotheses $h, h' \in \mathcal{H}$,*

$$|\epsilon_S(h,h') - \epsilon_T(h,h')| \leq \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T).$$

Lemma 2. *(From [2, 8]) For any labeling functions f_1, f_2, and f_3, we have*

$$\epsilon(f_1,f_2) \leq \epsilon(f_1,f_3) + \epsilon(f_2,f_3).$$

B.1 VC-dimension bounds

Lemma 3. *(From Ben-David et al. [3]) Let $\mathcal{H}$ be a hypothesis space on $\mathcal{Z}$ with VC-dimension d. If $\mathcal{U}$ and $\mathcal{U}'$ are samples of size m from $\mathcal{D}$ and $\mathcal{D}'$ respectively and $\hat{d}_{\mathcal{H}}(\mathcal{U},\mathcal{U}')$ is the empirical $\mathcal{H}$-divergence between samples, then for any $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$d_{\mathcal{H}}(\mathcal{D},\mathcal{D}') \leq \hat{d}_{\mathcal{H}}(\mathcal{U},\mathcal{U}') + 4\sqrt{\frac{d\log(2m) + \log(\frac{2}{\delta})}{m}}.$$

Theorem 1. *Let $\mathcal{H}$ be a hypothesis space of VC dimension d. If $\mathcal{U}_{S_0^0}$, $\mathcal{U}_{S_1^0}$, $\mathcal{U}_{T_1^0}$, $\mathcal{U}_{T_0^0}$ are samples of size m' each, drawn from $\mathcal{D}_{S_0^0}$, $\mathcal{D}_{S_1^0}$, $\mathcal{D}_{T_0^0}$, and $\mathcal{D}_{T_1^0}$ respectively, then for any $\delta \in (0,1)$, with probability at least $1 - \delta$ (over the choice of samples), for every $g \in \mathcal{H}$ (where $\mathcal{H}$ is a symmetric hypothesis space) the distance from equal opportunity in the target space is bounded by*

$$\Delta_{EOp_T}(g) \leq \Delta_{EOp_S}(g) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^0},\mathcal{U}_{S_0^0}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^0},\mathcal{U}_{S_1^0})$$
$$+ 8\sqrt{\frac{2d\log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda_0^0 + \lambda_1^0,$$

where $\lambda_\alpha^l = \epsilon_{S_\alpha^l}(g^*,f) + \epsilon_{T_\alpha^l}(g^*,f)$.

Proof. Without loss of generality assume $\mathbb{E}_{Z_0^0 \sim D_{S_0^0}} \geq \mathbb{E}_{Z_1^0 \sim D_{S_1^0}}$. Then we can rewrite $\Delta_{EOp_S}(g)$ as follows:

$$
\begin{aligned}
\Delta_{EOp_S}(g) &= \mathbb{E}_{Z_0^0 \sim \mathcal{D}_{S_0^0}}\left[g(Z_0^0)\right] - \mathbb{E}_{Z_1^0 \sim \mathcal{D}_{S_1^0}}\left[g(z_1^0)\right] \\
&= \mathbb{E}_{Z_0^0 \sim \mathcal{D}_{S_0^0}}\left[g(Z_0^0)\right] + \mathbb{E}_{Z_1^0 \sim \mathcal{D}_{S_1^0}}\left[1 - g(z_1^0)\right] - 1 \\
&= \epsilon_{S_0^0}(g, f) + \epsilon_{S_1^0}(1 - g, f) - 1,
\end{aligned}
$$

where the last line follows from the fact that equal opportunity only cares about the error on the false data-points.

We now have the tools to find an upper-bound on $\Delta_{EOp_T}(g)$.

$$
\begin{aligned}
\Delta_{EOp_T}(g) =&\, \epsilon_{T_0^0}(g, f) + \epsilon_{T_1^0}(1 - g, f) - 1 \\
\leq&\, \epsilon_{T_0^0}(g, g^*) + \epsilon_{T_0^0}(f, g^*) + \epsilon_{T_1^0}(1 - g, g^*) + \epsilon_{T_1^0}(f, g^*) - 1 \qquad (7) \\
=&\, \epsilon_{T_0^0}(g^*, f) + \epsilon_{T_0^0}(g, g^*) + \epsilon_{T_1^0}(g^*, f) + \epsilon_{T_1^0}(1 - g, g^*) - 1 \\
=&\, \epsilon_{T_0^0}(g^*, f) + \epsilon_{T_0^0}(g, g^*) + \epsilon_{S_0^0}(g, g^*) - \epsilon_{S_0^0}(g, g^*) \\
&+ \epsilon_{T_1^0}(g^*, f) + \epsilon_{T_1^0}(1 - g, g^*) + \epsilon_{S_1^0}(1 - g, g^*) - \epsilon_{S_1^0}(1 - g, g^*) - 1 \\
\leq&\, \epsilon_{T_0^0}(g^*, f) + \epsilon_{S_0^0}(g, g^*) + \left|\epsilon_{T_0^0}(g, g^*) - \epsilon_{S_0^0}(g, g^*)\right| \\
&+ \epsilon_{T_1^0}(g^*, f) + \epsilon_{S_1^0}(1 - g, g^*) + \left|\epsilon_{T_1^0}(1 - g, g^*) - \epsilon_{S_1^0}(1 - g, g^*)\right| - 1 \\
\leq&\, \epsilon_{T_0^0}(g^*, f) + \epsilon_{S_0^0}(g, g^*) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^0}, D_{S_0^0}) \\
&+ \epsilon_{T_1^0}(g^*, f) + \epsilon_{S_1^0}(1 - g, g^*) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^0}, D_{S_1^0}) - 1 \qquad (8) \\
\leq&\, \epsilon_{T_0^0}(g^*, f) + \epsilon_{S_0^0}(g, f) + \epsilon_{S_0^0}(g^*, f) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^0}, D_{S_0^0}) \\
&+ \epsilon_{T_1^0}(g^*, f) + \epsilon_{S_1^0}(1 - g, f) + \epsilon_{S_1^0}(g^*, f) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^0}, D_{S_1^0}) - 1 \qquad (9) \\
=&\, \epsilon_{S_0^0}(g, f) + \epsilon_{T_0^0}(g^*, f) + \epsilon_{S_0^0}(g^*, f) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^0}, D_{S_0^0}) \\
&+ \epsilon_{S_1^0}(1 - g, f) + \epsilon_{T_1^0}(g^*, f) + \epsilon_{S_1^0}(g^*, f) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^0}, D_{S_1^0}) - 1 \\
=&\, \epsilon_{S_0^0}(g, f) + \epsilon_{S_1^0}(1 - g, f) - 1 + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^0}, D_{S_0^0}) \\
&+ \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^0}, D_{S_1^0}) + \lambda_0^0 + \lambda_1^0 \qquad (10) \\
=&\, \Delta_{EOp_S}(g) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^0}, D_{S_0^0}) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^0}, D_{S_1^0}) + \lambda_0^0 + \lambda_1^0 \\
\leq&\, \Delta_{EOp_S}(g) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^0}, \mathcal{U}_{S_0^0}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^0}, \mathcal{U}_{S_1^0}) \\
&+ 8\sqrt{\frac{2d\log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda_0^0 + \lambda_1^0, \qquad (11)
\end{aligned}
$$

Where inequality 7 is due to lemma 2, inequality 8 is due to lemma 1 and the fact that $\mathcal{H}$ is a symmetric hypothesis space, inequality 9 is due to lemma 2, equality 10 is due to the definition of λ_α^l, and inequality 11 is due to lemma 3. $\qquad\square$

Theorem 2. *Let $\mathcal{H}$ be a hypothesis space of VC dimension d. If $\mathcal{U}_{S_\alpha^l}$ are samples of size m' each, drawn from $\mathcal{D}_{S_\alpha^l}$ for all $\alpha \in \Omega_A = \{0, 1\}$ and $l \in \Omega_{\mathcal{Y}} = 0, 1$, then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ (over the choice of samples), for every $g \in \mathcal{H}$ (where $\mathcal{H}$ is a symmetric hypothesis space) the distance from equalized odds in the target space is bounded by*

$$
\Delta_{EO_T}(g) \leq \Delta_{EO_S}(g) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^0}, \mathcal{U}_{S_0^0}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^0}, \mathcal{U}_{S_1^0})
$$

$$+ \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^1}, \mathcal{U}_{S_0^1}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^1}, \mathcal{U}_{S_1^1})$$

$$+ 16\sqrt{\frac{2d\log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda_{EO},$$

where $\lambda_{EO} = \lambda_0^0 + \lambda_1^0 + \lambda_0^1 + \lambda_1^1$, and $\lambda_\alpha^l = \epsilon_{S_\alpha^l}(g^*, f) + \epsilon_{T_\alpha^l}(g^*, f)$.

Proof. WLOG assume $\mathbb{E}_{Z_0^0 \sim D_{S_0^0}}[g] \geq \mathbb{E}_{Z_1^0 \sim D_{S_1^0}}[g]$ and $\mathbb{E}_{Z_0^1 \sim D_{S_0^1}}[g] \geq \mathbb{E}_{Z_1^1 \sim D_{S_1^1}}[g]$. Then,

$$\Delta_{EO_S} = \mathbb{E}_{Z_0^0 \sim D_{S_0^0}}[g] - \mathbb{E}_{Z_1^0 \sim D_{S_1^0}}[g] + \mathbb{E}_{Z_0^1 \sim D_{S_0^1}}[g] - \mathbb{E}_{Z_1^1 \sim D_{S_1^1}}[g]$$

$$= \mathbb{E}_{Z_0^0 \sim D_{S_0^0}}[g] + \mathbb{E}_{Z_1^0 \sim D_{S_1^0}}[1-g] + \mathbb{E}_{Z_0^1 \sim D_{S_0^1}}[g] + \mathbb{E}_{Z_1^1 \sim D_{S_1^1}}[1-g] - 2$$

$$= \epsilon_{S_0^0}(g, f) + \epsilon_{S_1^0}(1-g, f) + \epsilon_{S_0^1}(g, f) + \epsilon_{S_1^1}(1-g, f) - 2.$$

Using this and the previous lemmas we have:

$$\Delta_{EO_T}(g) = \epsilon_{T_0^0}(g, f) + \epsilon_{T_1^0}(1-g, f) + \epsilon_{T_0^1}(g, f) + \epsilon_{T_1^1}(1-g, f) - 2$$

$$\leq \epsilon_{T_0^0}(g, g^*) + \epsilon_{T_0^0}(f, g^*) + \epsilon_{T_1^0}(1-g, g^*) + \epsilon_{T_1^0}(f, g^*)$$

$$\qquad + \epsilon_{T_0^1}(g, g^*) + \epsilon_{T_0^1}(f, g^*) + \epsilon_{T_1^1}(1-g, g^*) + \epsilon_{T_1^1}(f, g^*) - 2 \qquad (12)$$

$$= \epsilon_{T_0^0}(g^*, f) + \epsilon_{T_0^0}(g, g^*) + \epsilon_{S_0^0}(g, g^*) - \epsilon_{S_0^0}(g, g^*)$$

$$\qquad + \epsilon_{T_1^0}(g^*, f) + \epsilon_{T_1^0}(1-g, g^*) + \epsilon_{S_1^0}(1-g, g^*) - \epsilon_{S_1^0}(1-g, g^*)$$

$$\qquad + \epsilon_{T_0^1}(g^*, f) + \epsilon_{T_0^1}(g, g^*) + \epsilon_{S_0^1}(g, g^*) - \epsilon_{S_0^1}(g, g^*)$$

$$\qquad + \epsilon_{T_1^1}(f, g^*) + \epsilon_{T_1^1}(1-g, g^*) + \epsilon_{S_1^1}(1-g, g^*) - \epsilon_{S_1^1}(1-g, g^*) - 2$$

$$\leq \epsilon_{T_0^0}(g^*, f) + \epsilon_{S_0^0}(g, g^*) + \left| \epsilon_{T_0^0}(g, g^*) - \epsilon_{S_0^0}(g, g^*) \right|$$

$$\qquad + \epsilon_{T_1^0}(g^*, f) + \epsilon_{S_1^0}(1-g, g^*) + \left| \epsilon_{T_1^0}(1-g, g^*) - \epsilon_{S_1^0}(1-g, g^*) \right|$$

$$\qquad + \epsilon_{T_0^1}(g^*, f) + \epsilon_{S_0^1}(g, g^*) + \left| \epsilon_{T_0^1}(g, g^*) - \epsilon_{S_0^1}(g, g^*) \right|$$

$$\qquad + \epsilon_{T_1^1}(f, g^*) + \epsilon_{S_1^1}(1-g, g^*) + \left| \epsilon_{T_1^1}(1-g, g^*) - \epsilon_{S_1^1}(1-g, g^*) \right| - 2$$

$$\leq \epsilon_{T_0^0}(g^*, f) + \epsilon_{S_0^0}(g, g^*) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^0}, D_{S_0^0})$$

$$\qquad + \epsilon_{T_1^0}(g^*, f) + \epsilon_{S_1^0}(1-g, g^*) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^0}, D_{S_1^0})$$

$$\qquad + \epsilon_{T_0^1}(g^*, f) + \epsilon_{S_0^1}(g, g^*) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^1}, D_{S_0^1})$$

$$\qquad + \epsilon_{T_1^1}(f, g^*) + \epsilon_{S_1^1}(1-g, g^*) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^1}, D_{S_1^1}) - 2 \qquad (13)$$

$$\leq \epsilon_{T_0^0}(g^*, f) + \epsilon_{S_0^0}(g, f) + \epsilon_{S_0^0}(g^*, f) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^0}, D_{S_0^0})$$

$$\qquad + \epsilon_{T_1^0}(g^*, f) + \epsilon_{S_1^0}(1-g, f) + \epsilon_{S_1^0}(g^*, f) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^0}, D_{S_1^0})$$

$$\qquad + \epsilon_{T_0^1}(g^*, f) + \epsilon_{S_0^1}(g, f) + \epsilon_{S_0^1}(g^*, f) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^1}, D_{S_0^1})$$

$$\qquad + \epsilon_{T_1^1}(f, g^*) + \epsilon_{S_1^1}(1-g, f) + \epsilon_{S_1^1}(g^*, f) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^1}, D_{S_1^1}) - 2 \qquad (14)$$

$$= \lambda_0^0 + \epsilon_{S_0^0}(g, f) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^0}, D_{S_0^0})$$

$$\qquad + \lambda_1^0 + \epsilon_{S_1^0}(1-g, f) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^0}, D_{S_1^0})$$

$$\qquad + \lambda_0^1 + \epsilon_{S_0^1}(g, f) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^1}, D_{S_0^1})$$

$$+ \lambda_1^1 + \epsilon_{S_1^1}(1-g,f) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^1}, D_{S_1^1}) - 2$$

$$= \Delta_{EO_S}(g) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^0}, D_{S_0^0}) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^0}, D_{S_1^0})$$

$$+ \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^1}, D_{S_0^1}) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^1}, D_{S_1^1}) + \lambda_{EO}$$

$$\leq \Delta_{EO_S}(g) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^0}, \mathcal{U}_{S_0^0}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^0}, \mathcal{U}_{S_1^0})$$

$$+ \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^1}, \mathcal{U}_{S_0^1}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^1}, \mathcal{U}_{S_1^1})$$

$$+ 16\sqrt{\frac{2d\log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda_{EO}, \tag{15}$$

where inequality 12 is due to lemma 2, inequality 13 is due to lemma 1 and the fact that $\mathcal{H}$ is a symmetric hypothesis space, inequality 14 is due to lemma 2, and inequality 15 is due to lemma 3. $\qquad\square$

B.2 Rademacher Complexity Bounds

Lemma 4. *(A modification of Corollary 7 from Mansour et al. [22]) Let $\mathcal{H}$ by a hypothesis set of classifiers mapping the feature space X to the labels $\{-1, 1\}$. Let $\mathcal{U}$ and $\mathcal{U}'$ be the set of samples each of size m sampled from $\mathcal{D}$ and $\mathcal{D}'$ respectively. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over samples $\mathcal{U}$ and $\mathcal{U}'$:*

$$d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') \leq \hat{d}_{\mathcal{H}}(\mathcal{U}, \mathcal{U}') + 4\left(\mathfrak{R}_{\mathcal{U}}(\mathcal{H}) + \mathfrak{R}_{\mathcal{U}}(\mathcal{H})\right) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

Theorem 3. *Let $\mathcal{H}$ be a hypothesis space. If $\mathcal{U}_{S_0^{-1}}$, $\mathcal{U}_{S_1^{-1}}$, $\mathcal{U}_{T_0^{-1}}$, $\mathcal{U}_{T_1^{-1}}$ are samples of size m' each, drawn from $\mathcal{D}_{S_0^{-1}}$, $\mathcal{D}_{S_1^{-1}}$, $\mathcal{D}_{T_0^{-1}}$, and $\mathcal{D}_{T_1^{-1}}$ respectively, then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ (over the choice of samples), for every $g \in \mathcal{H}$ (where $\mathcal{H}$ is a symmetric hypothesis space) the distance from equal opportunity in the target space is bounded by*

$$\Delta_{EOp_T}(g) \leq \Delta_{EOp_S}(g) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^{-1}}, \mathcal{U}_{S_0^{-1}}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^{-1}}, \mathcal{U}_{S_1^{-1}})$$

$$+ 2\left(\mathfrak{R}_{U_{T_0^{-1}}}(\mathcal{H}) + \mathfrak{R}_{U_{S_0^{-1}}}(\mathcal{H}) + \mathfrak{R}_{U_{T_1^{-1}}}(\mathcal{H}) + \mathfrak{R}_{U_{S_1^{-1}}}(\mathcal{H})\right)$$

$$+ 6\sqrt{\frac{\log\frac{2}{\delta}}{2m}} + \lambda_0^{-1} + \lambda_1^{-1},$$

where $\lambda_\alpha^l = \epsilon_{S_\alpha^l}(g^*, f) + \epsilon_{T_\alpha^l}(g^*, f)$.

Proof. Without loss of generality assume $\mathbb{E}_{Z_0^{-1} \sim D_{S_0^{-1}}} \geq \mathbb{E}_{Z_1^{-1} \sim D_{S_1^{-1}}}$. Then we can rewrite Δ_{EOp_S} as follows.

$$\Delta_{EOp_S}(g) = \mathbb{E}_{Z_0^{-1} \sim D_{S_0^{-1}}}\left[\frac{1 + g(z_0^{-1})}{2}\right] - \mathbb{E}_{Z_1^{-1} \sim D_{S_1^{-1}}}\left[\frac{1 + g(z_1^{-1})}{2}\right]$$

$$= \mathbb{E}_{Z_0^{-1} \sim D_{S_0^{-1}}}\left[\frac{1 + g(z_0^{-1})}{2}\right] + \mathbb{E}_{Z_1^{-1} \sim D_{S_1^{-1}}}\left[1 - \frac{1 + g(z_1^{-1})}{2}\right] - 1$$

$$= \mathbb{E}_{Z_0^{-1} \sim D_{S_0^{-1}}}\left[\frac{1 + g(z_0^{-1})}{2}\right] + \mathbb{E}_{Z_1^{-1} \sim D_{S_1^{-1}}}\left[\frac{1 - g(z_1^{-1})}{2}\right] - 1$$

$$= \mathbb{E}_{Z_0^{-1} \sim D_{S_0^{-1}}}\left[\frac{g(z_0^{-1}) - f(z_0^{-1})}{2}\right] + \mathbb{E}_{Z_1^{-1} \sim D_{S_1^{-1}}}\left[\frac{g(z_1^{-1}) + f(z_1^{-1})}{2}\right] - 1 \tag{16}$$

$$= \epsilon_{S_0^{-1}}(g, f) + \epsilon_{S_1^{-1}}(-g, f) - 1,$$

where 16 is due to the fact that $f(z_0^{-1}) = -1$ by definition.

We now have the tools to find an upper bound on $\Delta_{EОp_T}(g)$.

$$
\begin{aligned}
\Delta_{EOp_T}(g) &= \epsilon_{T_0^{-1}}(g, f) + \epsilon_{T_1^{-1}}(-g, f) - 1 \\
&\leq \epsilon_{T_0^{-1}}(g, g^*) + \epsilon_{T_0^{-1}}(f, g^*) + \epsilon_{T_1^{-1}}(-g, g^*) + \epsilon_{T_1^{-1}}(f, g^*) - 1 \quad (17) \\
&= \epsilon_{T_0^{-1}}(f, g^*) + \epsilon_{T_0^{-1}}(g, g^*) + \epsilon_{S_0^{-1}}(g, g^*) - \epsilon_{S_0^{-1}}(g, g^*) \\
&\quad + \epsilon_{T_1^{-1}}(f, g^*) + \epsilon_{T_1^{-1}}(-g, g^*) + \epsilon_{S_1^{-1}}(-g, g^*) - \epsilon_{S_1^{-1}}(-g, g^*) - 1 \\
&\leq \epsilon_{T_0^{-1}}(g^*, f) + \epsilon_{S_0^{-1}}(g, g^*) + |\epsilon_{T_0^{-1}}(g, g^*) - \epsilon_{S_0^{-1}}(g, g^*)| \\
&\quad + \epsilon_{T_1^{-1}}(g^*, f) + \epsilon_{S_1^{-1}}(-g, g^*) + |\epsilon_{T_1^{-1}}(-g, g^*) - \epsilon_{S_1^{-1}}(-g, g^*)| - 1 \\
&\leq \epsilon_{T_0^{-1}}(g^*, f) + \epsilon_{S_0^{-1}}(g, g^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}}) \\
&\quad + \epsilon_{T_1^{-1}}(g^*, f) + \epsilon_{S_1^{-1}}(-g, g^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}}) - 1 \quad (18) \\
&\leq \epsilon_{T_0^{-1}}(g^*, f) + \epsilon_{S_0^{-1}}(g, f) + \epsilon_{S_0^{-1}}(g^*, f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}}) \\
&\quad + \epsilon_{T_1^{-1}}(g^*, f) + \epsilon_{S_1^{-1}}(-g, f) + \epsilon_{S_1^{-1}}(g^*, f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}}) - 1 \quad (19) \\
&= \epsilon_{S_0^{-1}}(g, f) + \epsilon_{S_1^{-1}}(-g, f) - 1 + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}}) \\
&\quad + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}}) + \lambda_0^{-1} + \lambda_1^{-1} \\
&= \Delta_{EOp_S}(g) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}}) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}}) + \lambda_0^{-1} + \lambda_1^{-1} \quad (20) \\
&\leq \Delta_{EOp_S}(g) + \lambda_0^{-1} + \lambda_1^{-1} \\
&\quad + \frac{1}{2} \left(\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}}) + 4 \left(\hat{\mathfrak{R}}_{U_{T_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{S_0^{-1}}}(\mathcal{H}) \right) + 6 \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \right) \\
&\quad + \frac{1}{2} \left(\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}}) + 4 \left(\hat{\mathfrak{R}}_{U_{T_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{S_0^{-1}}}(\mathcal{H}) \right) + 6 \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \right) \quad (21) \\
&= \Delta_{EOp_S}(g) + \frac{1}{2} \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}}) + \frac{1}{2} \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}}) \\
&\quad + 2 \left(\hat{\mathfrak{R}}_{U_{T_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{S_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{S_0^{-1}}}(\mathcal{H}) \right) \\
&\quad + 6 \sqrt{\frac{\log \frac{2}{\delta}}{2m}} + \lambda_0^{-1} + \lambda_1^{-1},
\end{aligned}
$$

where Eq. 17 is due to Lemma 2, Eq. 18 is due to Lemma 1, Eq. 19 is due to Lemma 2, Eq. 20 is due to the definition of $\Delta_{EOp_S}(g)$, and Eq. 21 is due to Lemma 4. $\qquad \square$

Theorem 4. *Let $\mathcal{H}$ be a hypothesis space. If $\mathcal{U}_{S_0^{-1}}$, $\mathcal{U}_{S_1^{-1}}$, $\mathcal{U}_{T_0^{-1}}$, $\mathcal{U}_{T_1^{-1}}$ $\mathcal{U}_{S_0^1}$, $\mathcal{U}_{S_1^1}$, $\mathcal{U}_{T_0^1}$, $\mathcal{U}_{T_1^1}$ are samples of size m' each, drawn from $\mathcal{D}_{S_0^{-1}}$, $\mathcal{D}_{S_1^{-1}}$, $\mathcal{D}_{T_0^{-1}}$, $\mathcal{D}_{T_1^{-1}}$, $\mathcal{D}_{S_0^1}$, $\mathcal{D}_{S_1^1}$, $\mathcal{D}_{T_0^1}$, and $\mathcal{D}_{T_1^1}$ respectively, then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ (over the choice of samples), for every $g \in \mathcal{H}$ (where $\mathcal{H}$ is a symmetric hypothesis space) the distance from equalized odds in the target space is bounded by*

$$
\begin{aligned}
\Delta_{EO_T}(g) \leq \Delta_{EO_S}(g) &+ \frac{1}{2} \Big(\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_0^{-1}}, \mathcal{U}_{T_0^{-1}}) + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_1^{-1}}, \mathcal{U}_{T_1^{-1}}) \\
&+ \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_0^1}, \mathcal{U}_{T_0^1}) + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_1^1}, \mathcal{U}_{T_1^1}) \Big) \\
&+ 2 \Big(\hat{\mathfrak{R}}_{U_{S_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{S_1^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_1^{-1}}}(\mathcal{H}) \\
&+ \hat{\mathfrak{R}}_{U_{S_0^1}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_0^1}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{S_1^1}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_1^1}}(\mathcal{H}) \Big)
\end{aligned}
$$

$$+ 12\sqrt{\frac{\log\frac{2}{\delta}}{2m}} + \lambda_{EO},$$

where $\lambda_{EO} = \lambda_0^{-1} + \lambda_1^{-1} + \lambda_0^1 + \lambda_1^1$, and $\lambda_\alpha^l = \epsilon_{S_\alpha^l}(g^*, f) + \epsilon_{T_\alpha^l}(g^*, f)$.

Proof. Without loss of generality assume $\mathbb{E}_{Z_0^{-1} \sim D_{S_0^{-1}}} \geq \mathbb{E}_{Z_1^{-1} \sim D_{S_1^{-1}}}$ and $\mathbb{E}_{Z_0^1 \sim D_{S_0^1}} \geq \mathbb{E}_{Z_1^1 \sim D_{S_1^1}}$.
Then we can rewrite $\Delta_{EO_{p_S}}$ as follows.

$$\Delta_{EO_T}(g) = \mathbb{E}_{Z_0^{-1} \sim D_{T_0^{-1}}} \left[\frac{1 + g(z_0^{-1})}{2} \right] - \mathbb{E}_{Z_1^{-1} \sim D_{T_1^{-1}}} \left[\frac{1 + g(z_1^{-1})}{2} \right]$$
$$+ \mathbb{E}_{Z_0^1 \sim D_{T_0^1}} \left[\frac{1 + g(z_0^1)}{2} \right] - \mathbb{E}_{Z_1^1 \sim D_{T_1^1}} \left[\frac{1 + g(z_1^1)}{2} \right]$$
$$= \mathbb{E}_{Z_0^{-1} \sim D_{T_0^{-1}}} \left[\frac{1 + g(z_0^{-1})}{2} \right] + \mathbb{E}_{Z_1^{-1} \sim D_{T_1^{-1}}} \left[1 - \frac{1 + g(z_1^{-1})}{2} \right] - 1$$
$$+ \mathbb{E}_{Z_0^1 \sim D_{T_0^1}} \left[\frac{1 + g(z_0^1)}{2} \right] + \mathbb{E}_{Z_1^1 \sim D_{T_1^1}} \left[1 - \frac{1 + g(z_1^1)}{2} \right] - 1$$
$$= \mathbb{E}_{Z_0^{-1} \sim D_{T_0^{-1}}} \left[\frac{1 + g(z_0^{-1})}{2} \right] + \mathbb{E}_{Z_1^{-1} \sim D_{T_1^{-1}}} \left[\frac{1 - g(z_1^{-1})}{2} \right]$$
$$+ \mathbb{E}_{Z_0^1 \sim D_{T_0^1}} \left[\frac{1 + g(z_0^1)}{2} \right] + \mathbb{E}_{Z_1^1 \sim D_{T_1^1}} \left[\frac{1 - g(z_1^1)}{2} \right] - 2$$
$$= \mathbb{E}_{Z_0^{-1} \sim D_{T_0^{-1}}} \left[\frac{|g(z_0^{-1}) - f(z_0^{-1})|}{2} \right] + \mathbb{E}_{Z_1^{-1} \sim D_{T_1^{-1}}} \left[\frac{|g(z_1^{-1}) + f(z_1^{-1})|}{2} \right]$$
$$+ \mathbb{E}_{Z_0^1 \sim D_{T_0^1}} \left[\frac{|g(z_0^1) - f(z_0^1)|}{2} \right] + \mathbb{E}_{Z_1^1 \sim D_{T_1^1}} \left[\frac{|g(z_1^1) + f(z_1^1)|}{2} \right] - 2$$
$$= \epsilon_{T_0^{-1}}(g, f) + \epsilon_{T_1^{-1}}(-g, f) + \epsilon_{T_0^1}(g, f) + \epsilon_{T_1^1}(-g, f) - 2$$

Using this and previous lemmas we have

$$\Delta_{EO_T}(g) = \epsilon_{T_0^{-1}}(g, f) + \epsilon_{T_1^{-1}}(-g, f) + \epsilon_{T_0^1}(g, f) + \epsilon_{T_1^1}(-g, f) - 2$$
$$\leq \epsilon_{T_0^{-1}}(g, g^*) + \epsilon_{T_0^{-1}}(f, g^*) + \epsilon_{T_1^{-1}}(-g, g^*) + \epsilon_{T_1^{-1}}(f, g^*)$$
$$+ \epsilon_{T_0^1}(g, g^*) + \epsilon_{T_0^1}(f, g^*) + \epsilon_{T_1^1}(-g, g^*) + \epsilon_{T_1^1}(f, g^*) - 2 \quad\quad (22)$$
$$= \epsilon_{T_0^{-1}}(f, g^*) + \epsilon_{T_0^{-1}}(g, g^*) + \epsilon_{S_0^{-1}}(g, g^*) - \epsilon_{S_0^{-1}}(g, g^*)$$
$$+ \epsilon_{T_1^{-1}}(f, g^*) + \epsilon_{T_1^{-1}}(-g, g^*) + \epsilon_{S_1^{-1}}(-g, g^*) - \epsilon_{S_1^{-1}}(-g, g^*)$$
$$+ \epsilon_{T_0^1}(f, g^*) + \epsilon_{T_0^1}(g, g^*) + \epsilon_{S_0^1}(g, g^*) - \epsilon_{S_0^1}(g, g^*)$$
$$+ \epsilon_{T_1^1}(f, g^*) + \epsilon_{T_1^1}(-g, g^*) + \epsilon_{S_1^1}(-g, g^*) - \epsilon_{S_1^1}(-g, g^*) - 2$$
$$\leq \epsilon_{T_0^{-1}}(f, g^*) + \epsilon_{S_0^{-1}}(g, g^*) + \left| \epsilon_{T_0^{-1}}(g, g^*) - \epsilon_{S_0^{-1}}(g, g^*) \right|$$
$$+ \epsilon_{T_1^{-1}}(f, g^*) + \epsilon_{S_1^{-1}}(-g, g^*) + \left| \epsilon_{T_1^{-1}}(-g, g^*) - \epsilon_{S_1^{-1}}(-g, g^*) \right|$$
$$+ \epsilon_{T_0^1}(f, g^*) + \epsilon_{S_0^1}(g, g^*) + \left| \epsilon_{T_0^1}(g, g^*) - \epsilon_{S_0^1}(g, g^*) \right|$$
$$+ \epsilon_{T_1^1}(f, g^*) + \epsilon_{S_1^1}(-g, g^*) + \left| \epsilon_{T_1^1}(-g, g^*) - \epsilon_{S_1^1}(-g, g^*) \right| - 2$$
$$\leq \epsilon_{T_0^{-1}}(f, g^*) + \epsilon_{S_0^{-1}}(g, g^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}})$$
$$+ \epsilon_{T_1^{-1}}(f, g^*) + \epsilon_{S_1^{-1}}(-g, g^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}})$$
$$+ \epsilon_{T_0^1}(f, g^*) + \epsilon_{S_0^1}(g, g^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^1}, D_{S_0^1})$$
$$+ \epsilon_{T_1^1}(f, g^*) + \epsilon_{S_1^1}(-g, g^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^1}, D_{S_1^1}) - 2 \quad\quad (23)$$

$$\leq \epsilon_{T_0^{-1}}(f, g^*) + \epsilon_{S_0^{-1}}(g, f) + \epsilon_{S_0^{-1}}(g^*, f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}})$$

$$+ \epsilon_{T_1^{-1}}(f, g^*) + \epsilon_{S_1^{-1}}(-g, f) + \epsilon_{S_1^{-1}}(g^*, f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}})$$

$$+ \epsilon_{T_0^{1}}(f, g^*) + \epsilon_{S_0^{1}}(g, f) + \epsilon_{S_0^{1}}(g^*, f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{1}}, D_{S_0^{1}})$$

$$+ \epsilon_{T_1^{1}}(f, g^*) + \epsilon_{S_1^{1}}(-g, f) + \epsilon_{S_1^{1}}(g^*, f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{1}}, D_{S_1^{1}}) - 2 \qquad (24)$$

$$= \epsilon_{S_0^{-1}}(g, f) + \epsilon_{S_1^{-1}}(-g, f) + \epsilon_{S_0^{1}}(g, f) + \epsilon_{S_1^{1}}(-g, f) - 2$$

$$+ \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}}) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}})$$

$$+ \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{1}}, D_{S_0^{1}}) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{1}}, D_{S_1^{1}}) + \lambda_0^{-1} + \lambda_1^{-1} + \lambda_0^{1} + \lambda_1^{1}$$

$$= \Delta_{EO_S}(g) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}}) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}})$$

$$+ \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{1}}, D_{S_0^{1}}) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{1}}, D_{S_1^{1}}) + \lambda_0^{-1} + \lambda_1^{-1} + \lambda_0^{1} + \lambda_1^{1}$$

$$\leq \Delta_{EO_S}(g) + \lambda_0^{-1} + \lambda_1^{-1} + \lambda_0^{1} + \lambda_1^{1}$$

$$+ \frac{1}{2} \left(\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}}) + 4 \left(\hat{\mathfrak{R}}_{U_{S_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_0^{-1}}}(\mathcal{H}) \right) + 6 \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \right)$$

$$+ \frac{1}{2} \left(\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}}) + 4 \left(\hat{\mathfrak{R}}_{U_{S_1^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_1^{-1}}}(\mathcal{H}) \right) + 6 \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \right)$$

$$+ \frac{1}{2} \left(\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{1}}, D_{S_0^{1}}) + 4 \left(\hat{\mathfrak{R}}_{U_{S_0^{1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_0^{1}}}(\mathcal{H}) \right) + 6 \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \right)$$

$$+ \frac{1}{2} \left(\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{1}}, D_{S_1^{1}}) + 4 \left(\hat{\mathfrak{R}}_{U_{S_1^{1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_1^{1}}}(\mathcal{H}) \right) + 6 \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \right) \qquad (25)$$

$$= \Delta_{EO_S}(g) + \frac{1}{2} \left(\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_0^{-1}}, \mathcal{U}_{T_0^{-1}}) + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_1^{-1}}, \mathcal{U}_{T_1^{-1}}) \right.$$

$$\left. + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_0^{1}}, \mathcal{U}_{T_0^{1}}) + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_1^{1}}, \mathcal{U}_{T_1^{1}}) \right)$$

$$+ 2 \left(\hat{\mathfrak{R}}_{U_{S_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{S_1^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_1^{-1}}}(\mathcal{H}) \right.$$

$$\left. + \hat{\mathfrak{R}}_{U_{S_0^{1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_0^{1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{S_1^{1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_1^{1}}}(\mathcal{H}) \right)$$

$$+ 12 \sqrt{\frac{\log \frac{2}{\delta}}{2m}} + \lambda_0^{-1} + \lambda_1^{-1} + \lambda_0^{1} + \lambda_1^{1},$$

where Eq. 22 is due to Lemma 2, Eq. 23 is due to Lemma 1, Eq. 24 is due to Lemma 2, and 25 is due to Lemma 4. □

C Experimental setup

For the UCI adult dataset we used all 14 features as provided in `https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names`. The original train/test split is used. For the COMPAS dataset we used the features provided in `https://github.com/propublica/compas-analysis/blob/master/compas-scores.csv`, and predict the risk of recidivism (decile_score) for each row.

We did 10-fold cross-validation and choose the hyperparameters with the best performance on the validation data. 64 dimension embedding is used for categorical features and 256 hidden units are
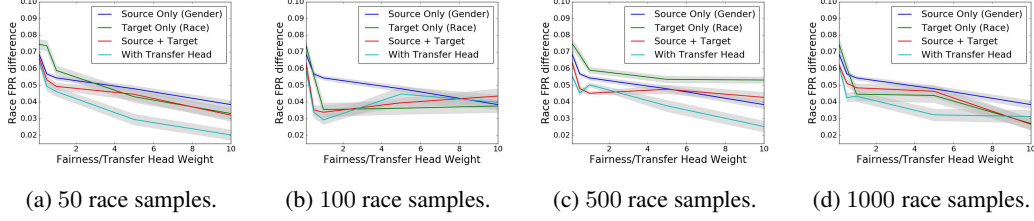
(a) 50 race samples.   (b) 100 race samples.   (c) 500 race samples.   (d) 1000 race samples.

Figure 5: Gender → Race on the UCI dataset. Comparison of FPR difference on sensitive attribute *race*, by transferring from the source domain (1000 samples for each gender) to the target domain (varying samples for each race as indicated in the caption).
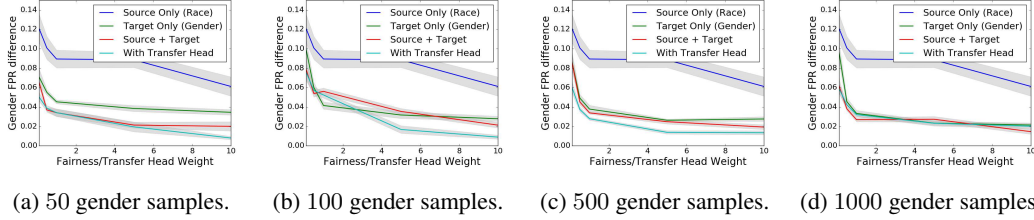

(a) 50 gender samples.   (b) 100 gender samples.   (c) 500 gender samples.   (d) 1000 gender samples.

Figure 6: Race → Gender on the UCI dataset. Comparison of FPR difference on sensitive attribute *gender*, by transferring from the source domain (1000 samples for each gender) to the target domain (varying samples for each race as indicated in the caption).

used in the model. We did parameter search and found 10K steps yields a good balance of runtime and accuracy. Each run takes about 1hr for UCI data and 0.5hrs for COMPAS on a single CPU with 2GB RAM. Increasing learning rate speeds up experiments but also hurts accuracy slightly (e.g., ~2pp decrease on UCI).

For range of parameters, we have considered the following: (1) batch size: $[64, 128, 256, 512]$; (2) learning rate: $[0.01, 0.1, 1.0]$; (3) number of hidden units: $[64, 128, 256, 512]$; (4) embedding dimension: $[32, 64, 128]$. (5) number of steps: $[5000, 10000, 20000, 50000]$.

D   Experiments

D.1   Experiment Results for fairness on UCI and COMPAS

Figure 5 depicts the results of the analysis for transferring from gender to race, while Figure 6 shows the results for transferring from race to gender, on the UCI dataset. Figure 7 and Figure 8 show the results on the COMPAS dataset. The line and the shaded areas show the mean and the standard error of the mean across 30 trials. These experiments show that the Transfer model is effective in decreasing the FPR gap in the target domain and is more sample efficient than previous methods.
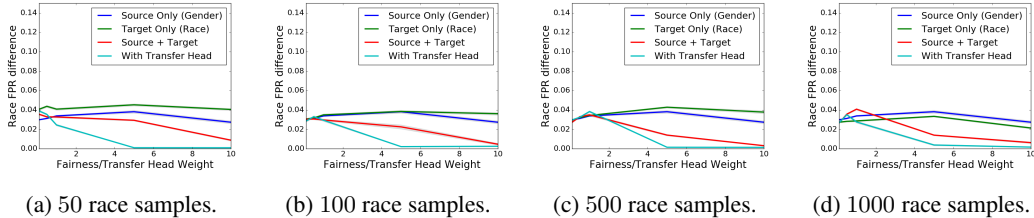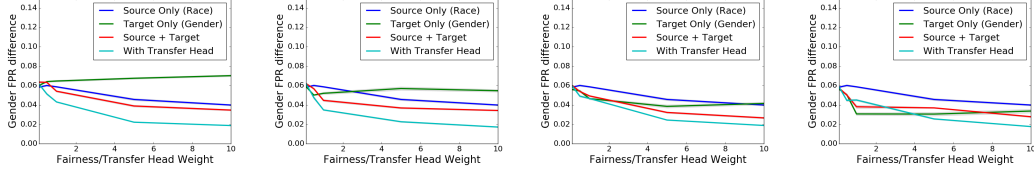

(a) 50 race samples.   (b) 100 race samples.   (c) 500 race samples.   (d) 1000 race samples.

Figure 7: Gender → Race on the COMPAS dataset. Comparison of FPR difference on sensitive attribute *race*, by transferring from the source domain (1000 samples for each gender) to the target domain (varying samples for each race as indicated in the caption).
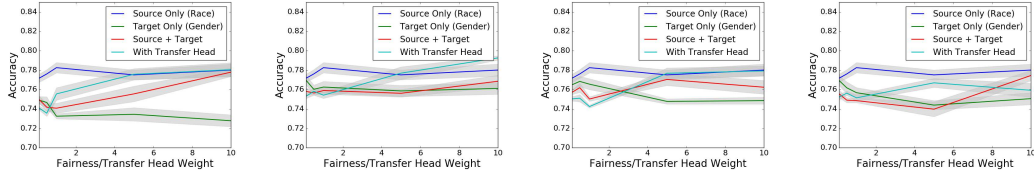
(a) 50 gender samples. (b) 100 gender samples. (c) 500 gender samples. (d) 1000 gender samples.

Figure 8: Race → Gender on the COMPAS dataset. Comparison of FPR difference on sensitive attribute *gender*, by transferring from the source domain (1000 samples for each gender) to the target domain (varying samples for each race as indicated in the caption).

D.2   Accuracy vs. Fairness/Transfer Head Weight

In this section we further add the comparison on accuracy with respect to the weight of the fairness/transfer head. Fig. 9 and Fig. 10 show the results comparing the Transfer model with the baselines, by transferring *race* to *gender*, and *race* to *gender*, respectively. Fig. 11 and Fig. 12 show the results on COMPAS.

(a) 50 gender samples. (b) 100 gender samples. (c) 500 gender samples. (d) 1000 gender samples.

Figure 9: Comparison of accuracy on the UCI data for Race → Gender, by transferring from the source domain (1000 samples for each race) to the target domain (varying samples for each gender as indicated in the caption).
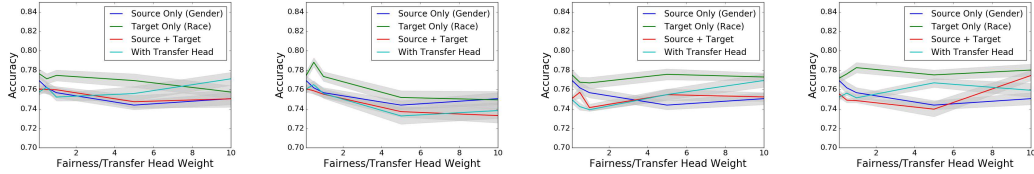
(a) 50 race samples. (b) 100 race samples. (c) 500 race samples. (d) 1000 race samples.

Figure 10: Comparison of accuracy on the UCI data for Gender → Race, by transferring from the source domain (1000 samples for each gender) to the target domain (varying samples for each race as indicated in the caption).
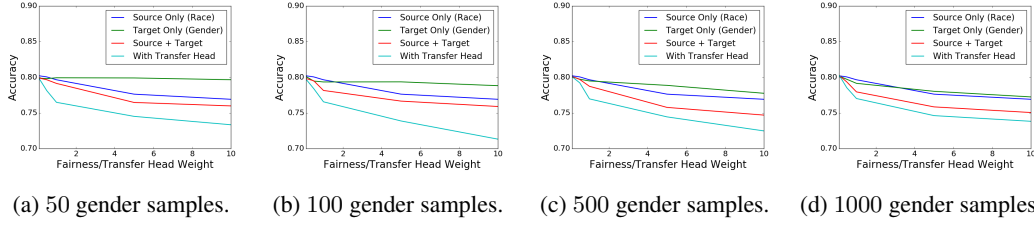
(a) 50 gender samples. (b) 100 gender samples. (c) 500 gender samples. (d) 1000 gender samples.

Figure 11: Comparison of accuracy on COMPAS for Race $\rightarrow$ Gender, by transferring from the source domain (1000 samples for each race) to the target domain (varying samples for each gender as indicated in the caption).
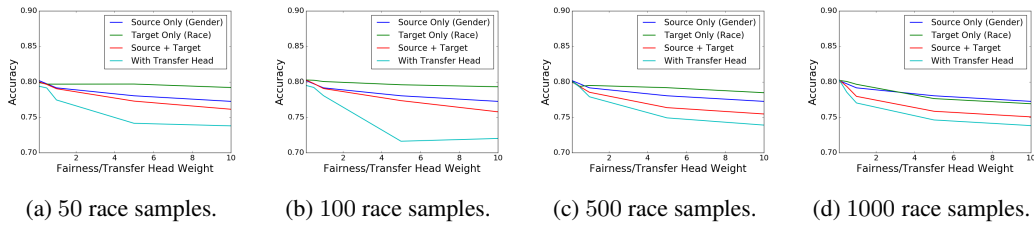

(a) 50 race samples. (b) 100 race samples. (c) 500 race samples. (d) 1000 race samples.

Figure 12: Comparison of accuracy on COMPAS for Gender $\rightarrow$ Race, by transferring from the source domain (1000 samples for each gender) to the target domain (varying samples for each race as indicated in the caption).