A Related Work

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

Domain Adaptation Both Pan et al. [26], and Weiss et al. [29] 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. [25] provide theoretical bounds on domain adaptation using Rademacher Complexity analysis. In later research, Ganin et al. [13] 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. [21] 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. [24] provides a theoretical framework for understanding how adversarial learning optimizes these fairness goals. Zhang et al. [31] use adversarial training over logits rather than hidden representations. Other work has focused on constraint-based optimization of fairness objectives [14, 1]. Tsipras et al. [28] 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. Kallus and Zhou [19] look at fairness in personalization when sensitive attributes are missing. Similarly, Chen et al. [8] look at measuring disparity when sensitive attributes are unknown.

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. Coston et al. [9] look at domain adaptation for fairness where sensitive attribute labels are not available in both the source and target domains. Kallus and Zhou [18] use covariate shift correction when computing fairness metrics to address bias in label collection. More related, Madras et al. [24] 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 [20] 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.

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, 10]) 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 $\mathcal{U}_{S_0^0}$, $\mathcal{U}_{S_1^0}$, $\mathcal{U}_{T_0^1}$, $\mathcal{U}_{T_1^0}$ 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 $\lambda_\alpha^l = \epsilon_{S_\alpha^l}(g^*, f) + \epsilon_{T_\alpha^l}(g^*, f)$.

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 && (4) \\
=& \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 && (5) \\
\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 && (6) \\
=& \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 && (7) \\
=& \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, && (8)
\end{aligned}$$

Where inequality 4 is due to lemma 2, inequality 5 is due to lemma 1 and the fact that $\mathcal{H}$ is a symmetric hypothesis space, inequality 6 is due to lemma 2, equality 7 is due to the definition of λ_α^l, and inequality 8 is due to lemma 3. □

C Experiment 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 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).

We considered the following range of parameters: (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: $[5000, 10000, 20000, 50000]$.

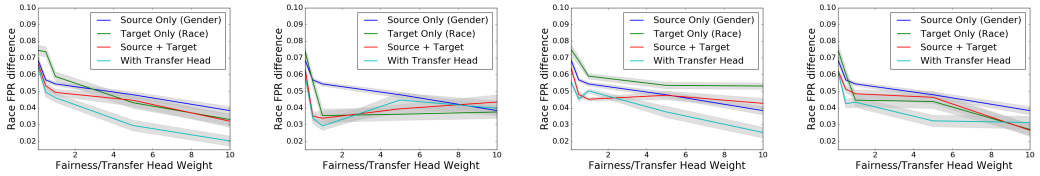
D Experiment Results

D.1 Experiment Results for fairness on UCI and COMPAS

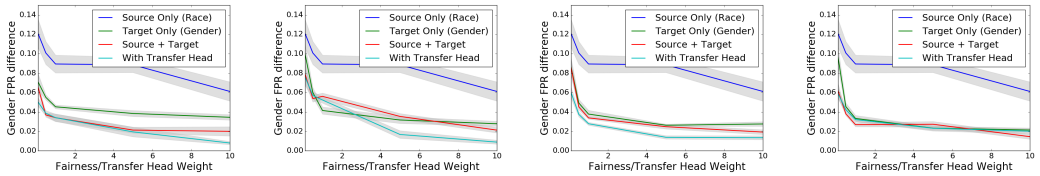
Figure 1 depicts the results of the analysis for transferring from gender to race, and from race to gender, respectively, on the UCI dataset. Figure 2 show the results on the COMPAS dataset. The line and the shaded areas show the mean and the standard error of the mean across 30 trials. These experiments show that the Transfer model is effective in decreasing the FPR gap in the target domain and is more sample efficient than previous methods.

D.2 Accuracy vs. Fairness/Transfer Head Weight

We further add the comparison on accuracy with respect to the weight of the fairness/transfer head. Fig. 3 show the results comparing the Transfer model with the baselines, by transferring *race* to *gender*, and *race* to *gender*, respectively, on the UCI dataset. Fig. 4 show the results on the COMPAS dataset.

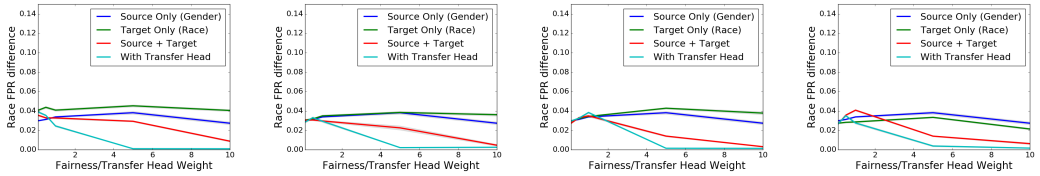


(a) 50 race samples. (b) 100 race samples. (c) 500 race samples. (d) 1000 race samples.

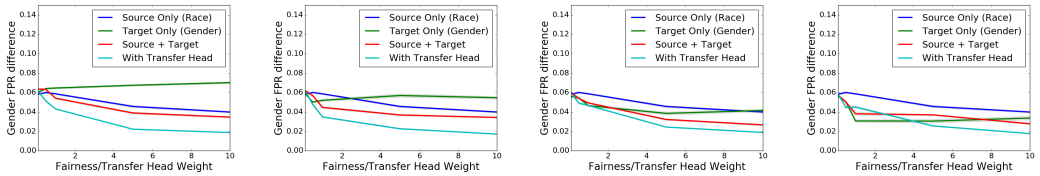


(e) 50 gender samples. (f) 100 gender samples. (g) 500 gender samples. (h) 1000 gender samples.

Figure 1: Transfer Gender to Race (first row) and Race to Gender (second row) on the UCI dataset. Comparison of FPR difference on the target sensitive attribute, by transferring from the source domain (1000 samples) to the target domain (varying samples as indicated in the caption).

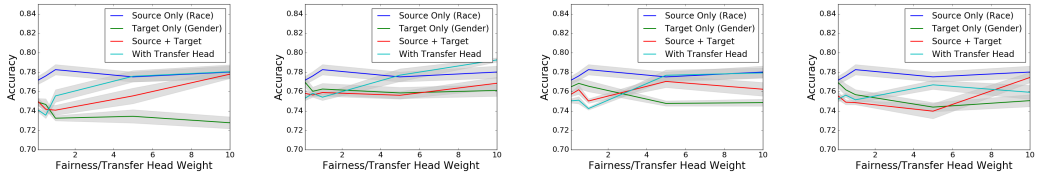


(a) 50 race samples. (b) 100 race samples. (c) 500 race samples. (d) 1000 race samples.

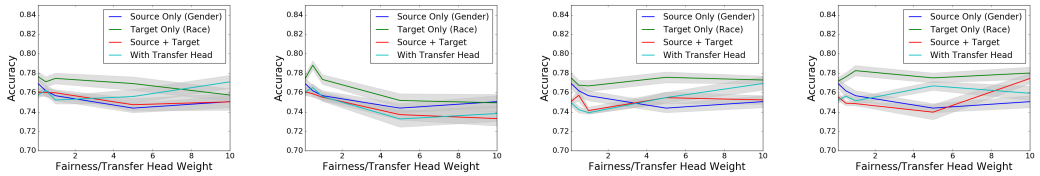


(e) 50 gender samples. (f) 100 gender samples. (g) 500 gender samples. (h) 1000 gender samples.

Figure 2: Transfer Gender to Race (first row) and Race to Gender (second row) on the COMPAS dataset. Comparison of FPR difference on the target sensitive attribute, by transferring from the source domain (1000 samples) to the target domain (varying samples as indicated in the caption).

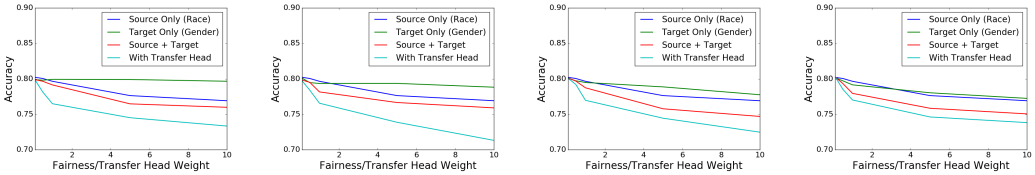


(a) 50 gender samples. (b) 100 gender samples. (c) 500 gender samples. (d) 1000 gender samples.

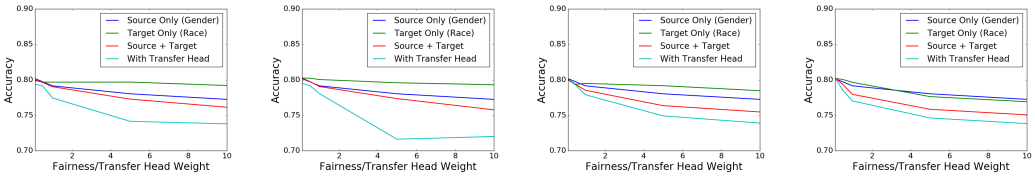


(e) 50 race samples. (f) 100 race samples. (g) 500 race samples. (h) 1000 race samples.

Figure 3: Comparison of accuracy on the UCI data for Race to Gender (first row), and Gender to Race (second row), by transferring from the source domain (1000 samples) to the target domain (varying samples as indicated in the caption).



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



(e) 50 race samples. (f) 100 race samples. (g) 500 race samples. (h) 1000 race samples.

Figure 4: Comparison of accuracy on COMPAS for Race to Gender (first row), and Gender to Race (second row), by transferring from the source domain (1000 samples) to the target domain (varying samples as indicated in the caption).