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

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 [27]. 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* [17] 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 [16], or researchers need very sample-efficient techniques [1, 4]. For distant proxies, researchers have asked how well fairness transfers across attributes [20]. 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

*Work done while at Google

AI for Social Good workshop at NeurIPS (2019), Vancouver, Canada.

[11]. 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.

2 Problem Formulation

We begin with some notation to make the problem formulation precise. 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., $\Pr_{Z\sim\mathcal{D}_S}[Z|A_S = \textit{male}] \neq \Pr_{Z\sim\mathcal{D}_S}[Z|A_S = \textit{female}]$), as well as a target domain $\mathcal{D}_T$ influenced by sensitive attribute $A_T \in \mathcal{A}_T$ (e.g., $\Pr_{Z\sim\mathcal{D}_T}[Z|A_T = \textit{black}] \neq \Pr_{Z\sim\mathcal{D}_T}[Z|A_T = \textit{white}]$). In order for this to be a domain adaptation problem, we assume $\Pr_{Z\sim\mathcal{D}_S}[Z|A_S] \neq \Pr_{Z\sim\mathcal{D}_T}[Z|A_T]$. 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 $\mathcal{H}$.

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 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 we consider a definition of *equality of opportunity* [17]. 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 $\Pr(\hat{Y} = 1|A = 0, Y = 0) = \Pr(\hat{Y} = 1|A = 1, Y = 0)$, 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).

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 $\Delta_{Fair_T}(g) \leq \Delta_{Fair_S}(g) + \epsilon$. The key question we hope to answer is: what is ϵ?

3 Bounds on Fairness in the Target Domain

To expand the key question 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

$$d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') = 2 \sup_{g\in\mathcal{H}} |\Pr_{\mathcal{D}}[I(g)] - \Pr_{\mathcal{D}'}[I(g)]|, \tag{1}$$

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

$$g \in \mathcal{H}\Delta\mathcal{H} \iff g(Z) = h(Z) \oplus h'(Z) \quad \text{for some } h, h' \in \mathcal{H}, \tag{2}$$

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*

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

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.

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 traditional domain adaptation, ignoring fairness, the entire domains should be close, 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. With 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.

4 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. Ganin et al. [13] proved that traditional domain adaptation can be framed as minimizing the distance between source and target with adversarial training. [23, 12, 4, 21] similarly have applied adversarial training to achieve fairness goals, and Madras et al. [24] proved that equality of odds can be optimized with adversarial training similar to domain adaptation. We build on this intuition to design a learning objective for transferring equality of opportunity to a target domain.

Recently, Zhang et al. [31] 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 [30, 5] minimizing the correlation between group and the one dimensional representation of the data. Like previous work, we find 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 [15, 22, 6] over predictions:

$$\min \Bigg[\sum_{Z \in \mathcal{D}_S \cup \mathcal{D}_T} L_Y(f(Z), g(Z)) + \sum_{(A, Z^0) \sim \mathcal{D}_{S^0}} \lambda_{Fair} L_{MMD}\left(a(h(Z^0)), A\right) + \sum_{(d, Z^0) \sim \left(\mathcal{D}_{S^0} \cup \mathcal{D}_{T^0}\right)} \lambda_{DA} L_{MMD}\left(d(h(Z^0)), d\right) \Bigg], \tag{3}$$

where $\lambda_{Fair} L_{MMD}\left(a(h(Z^0)), A\right)$ is the MMD regularization over the sensitive attributes in the source domain, and $\lambda_{DA} L_{MMD}\left(d(h(Z^0)), d\right)$ is the MMD regularization over source/target membership.

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, or balanced data – equally representing all necessary subgroups – should be used as in [24] and Eq. 3.

5 Experiments

We explore how and when our proposed modeling approach in Section 4 facilitates the transfer of fairness from the source to the target domain on two real-world datasets (The UCI Adult[2] and ProPublica's COMPAS recidivism data[3]). Note, we use these datasets for understanding our theory and model, and *not* as a comment on when or if the proposed tasks are appropriate, as in [1].

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

Effect of Target Sample Size 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: First, we vary the number of samples for each sensitive group in the target domain ($\{50, 100, 500, 1000\}$). 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 1 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.

			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	$\mathbf{0.027 \pm 0.018}$	$\mathbf{0.027 \pm 0.019}$	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 1: 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.

6 Conclusion

In this paper we provide the first theoretical examination of transfer of machine learning fairness across domains. We have provided theoretical bounds on the transfer of fairness for equal opportunity and, based on this theory, we developed a new modeling approach to transfer fairness to 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.

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

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

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] J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *FAT**, pages 339–348. ACM, 2019.

[9] A. Coston, K. N. Ramamurthy, D. Wei, K. R. Varshney, S. Speakman, Z. Mustahsan, and S. Chakraborty. Fair transfer learning with missing protected attributes. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, Honolulu, HI, USA*, 2019.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] N. Kallus and A. Zhou. Assessing disparate impacts of personalized interventions: Identifiability and bounds. *arXiv preprint arXiv:1906.01552*, 2019.

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

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

[22] 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.

[23] 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.

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

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

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

[27] 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.

[28] 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.

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

[30] 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.

[31] 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 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: $[5000, 10000, 20000, 50000]$.

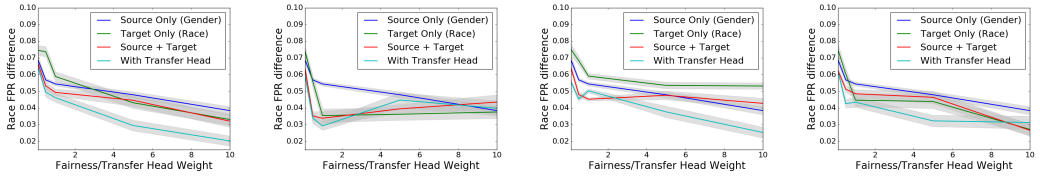
D Experiment Results

D.1 Experiment Results for fairness on UCI and COMPAS

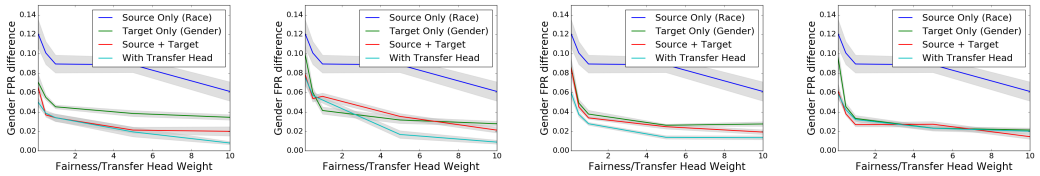
Figure 1 depicts the results of the analysis for transferring from gender to race, and from race to gender, respectively, on the UCI dataset. Figure 2 show the results on the COMPAS dataset. The line and the shaded areas show the mean and the standard error of the mean across 30 trials. These experiments 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.

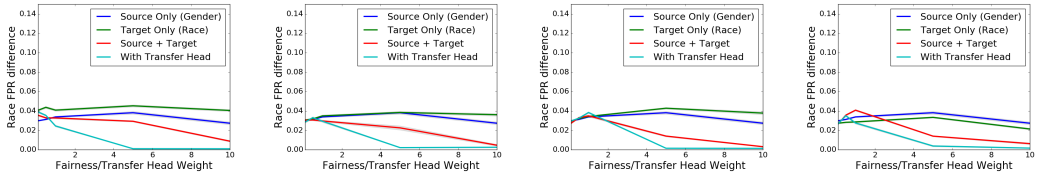


(a) 50 race samples. (b) 100 race samples. (c) 500 race samples. (d) 1000 race samples.

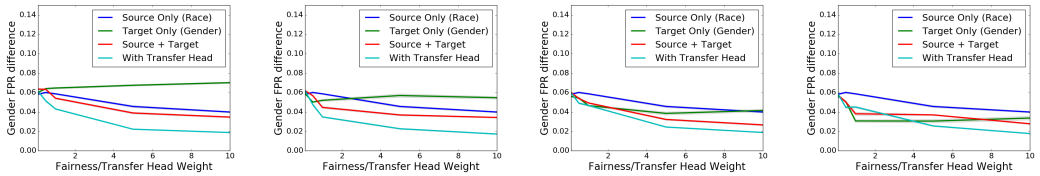


(e) 50 gender samples. (f) 100 gender samples. (g) 500 gender samples. (h) 1000 gender samples.

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).

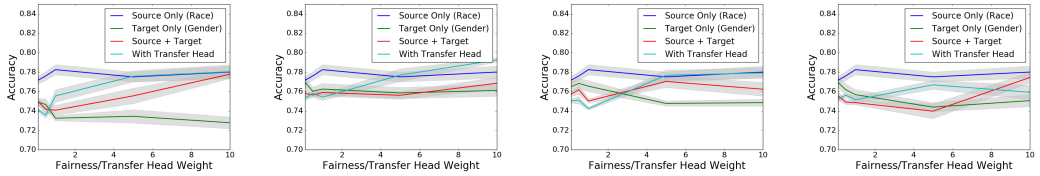


(a) 50 race samples. (b) 100 race samples. (c) 500 race samples. (d) 1000 race samples.

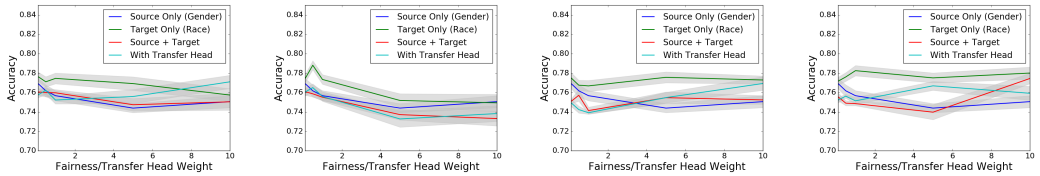


(e) 50 gender samples. (f) 100 gender samples. (g) 500 gender samples. (h) 1000 gender samples.

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).

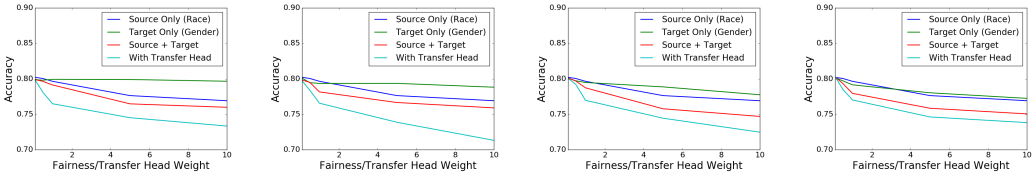


(a) 50 gender samples. (b) 100 gender samples. (c) 500 gender samples. (d) 1000 gender samples.

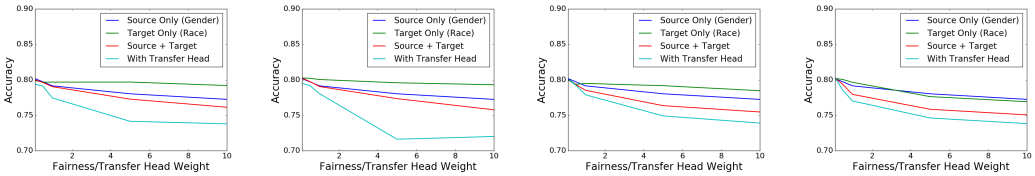


(e) 50 race samples. (f) 100 race samples. (g) 500 race samples. (h) 1000 race samples.

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).