Fairness under Covariate Shift: Improving Fairness-Accuracy Tradeoff with Few Unlabeled Test Samples

Shreyas Havaldar[*1] **, Jatin Chauhan**[*†2]**, Karthikeyan Shanmugam**[*1]**,**
Jay Nandy[†3]**, Aravindan Raghuveer**[1]

[1]Google Research India
[2]UCLA
[3]Fujitsu Research India
{shreyasjh, karthikeyanvs, araghuveer}@google.com, {chauhanjatin100, jayjaynandy}@gmail.com,

Abstract

Covariate shift in the test data is a common practical phenomena that can significantly downgrade both the accuracy and the fairness performance of the model. Ensuring fairness across different sensitive groups under covariate shift is of paramount importance due to societal implications like criminal justice. We operate in the unsupervised regime where only a small set of unlabeled test samples along with a labeled training set is available. Towards improving fairness under this highly challenging yet realistic scenario, we make three contributions. First is a novel composite weighted entropy based objective for prediction accuracy which is optimized along with a representation matching loss for fairness. We experimentally verify that optimizing with our loss formulation outperforms a number of state-of-the-art baselines in the pareto sense with respect to the fairness-accuracy tradeoff on several standard datasets. Our second contribution is a new setting we term Asymmetric Covariate Shift that, to the best of our knowledge, has not been studied before. Asymmetric covariate shift occurs when distribution of covariates of one group shifts significantly compared to the other groups and this happens when a dominant group is over-represented. While this setting is extremely challenging for current baselines, We show that our proposed method significantly outperforms them. Our third contribution is theoretical, where we show that our weighted entropy term along with prediction loss on the training set approximates test loss under covariate shift. Empirically and through formal sample complexity bounds, we show that this approximation to the unseen test loss does not depend on importance sampling variance which affects many other baselines.

1 Introduction

Predictions of machine learnt models are used to make important decisions that have societal impact, like in criminal justice, loan approvals, to name a few. Therefore, there is a lot of interest in understanding, analyzing and improving model performance along other dimensions like robustness (Silva and Najafirad 2020), model generalization (Wiles et al. 2021) and fairness (Oneto and Chiappa 2020). In this work, we focus on the algorithmic fairness aspect. Datasets used for training could be biased in the sense that some groups may be under-represented, thus biasing classifier decisions towards the over-represented group or the bias could be in terms of undesirable causal pathways between sensitive attribute and the label in the real world data generating mechanism (Oneto and Chiappa 2020). It has often been observed (Bolukbasi et al. 2016), (Buolamwini and Gebru 2018) that algorithms that optimize predictive accuracy that are fed pre-existing biases further learn and then propagate the same biases. Improving fairness of learning models has received significant attention from the research community (Mitchell et al. 2021).

Another common challenge that models deployed in real world situations face is that of *Covariate Shift*. In covariate shift, the distribution of covariates (feature vectors) across training and testing changes, however the optimal label predictor conditioned on input remains the same. Therefore the model may make wrong predictions when deployed or more seriously can slowly degrade over time when the covariate shift is gradual. Due to the practical importance of this problem, there has been a significant amount of research in detecting covariate shift and modeling methodologies to address it (Wilson and Cook 2020; Redko et al. 2020).

The problem that we study in this paper is at the juncture of the above two hard problems: ensuring fairness under covariate shift. While this question has not received much attention, some recent works like (Rezaei et al. 2021) have begun to address this problem. We also introduce a new variant of covariate shift called *Asymmetric covariate shift* where distribution of covariates of one group shifts significantly as compared to the other groups. Asymmetric covariate shift is a very common practical situation when there is long tail of underrepresented groups in the training data. For example, consider the popular and important task of click through prediction of advertisements (Li et al. 2015). Small and medium sized advertisers have poorer representation in the training data because they do not spend as much as the large businesses on advertising. Therefore during inference Small and Medium (SMB) advertisement clicks will see significantly more co-variate shift compared to those clicks on ads from large advertisers. Also, due to the nature of the problem of co-

[*]These authors contributed equally.
[†]Contributions to this work were made when affiliated with Google Research India

variate shift, access to large labeled test is often not possible. In summary, the problem we aim to tackle is "Provide a high fairness-accuracy tradeoff under both symmetric and asymmetric covariate shift while having acesss to a very small set of unlabeled test samples". To this end, we make three key contributions in this paper.

1. We introduce a composite objective to approximate the prediction loss on the unlabeled test that involves *a novel weighted entropy objective on the set of unlabeled test samples* along with ERM objective on the labeled training samples. We optimize these weights using *min-max* optimization that implicitly drives these weights to importance sampling ratios with no density estimation steps. We show that our proposed objective has *provably* lower variance compared to the importance sampling based methods. This composite objective is then combined with a representation matching loss to train fair classifiers. (Section 4).
2. We introduce a new type of covariate shift called *asymmetric covariate shift* wherein one protected group exhibits large covariate shift while the other does not. We highlight that fairness-accuracy tradeoff degrades under this case for existing methods (Section 3.3). We show empirically that the combination of our objective and representation matching achieves the best accuracy fariness-tradeoff even in this case.
3. By incorporating our proposed weighted entropy objective with the Wasserstein based representation matching across sub-groups, we empirically compare against a number of baseline methods on benchmark datasets. In particular, we achieve the best accuracy-equalized odds tradeoff in the *pareto sense*.

2 Related Work

Techniques for imposing fairness: *Pre-processing* techniques aim to transform the dataset (Calmon et al. 2017; Swersky, Pitassi, and Dwork 2013; Kamiran and Calders 2012) followed by a standard training. *In-processing* methods directly modify the learning algorithms using techniques, such as, adversarial learning (Madras et al. 2018; Zhang, Lemoine, and Mitchell 2018), (Agarwal et al. 2018; Cotter et al. 2019; Donini et al. 2018; Fish, Kun, and Lelkes 2016; Zafar et al. 2017; Celis et al. 2019). *Post-processing* approaches, primarily focus on modifying the outcomes of the predictive models in order to make unbiased predictions (Pleiss et al. 2017; Hardt, Price, and Srebro 2016)

Distribution Shift: Research addressing distribution shift in machine learning is vast and is growing. The general case considers a joint distribution shift between training and testing data (Ben-David et al. 2006; Blitzer et al. 2007; Moreno-Torres et al. 2012) resulting in techniques like domain adaptation (Ganin and Lempitsky 2015), distributionally robust optimization (Sagawa et al. 2019; Duchi and Namkoong 2021) and invariant risk minimization and its variants (Arjovsky et al. 2019; Shi et al. 2021). A survey of various methods and their relative performance is discussed by (Wiles et al. 2021). We focus on the problem of *Covariate Shift* where the *Conditional Label* distribution is invariant while there is a shift in the marginal distribution of the covariates across training and test samples. This classical setup is studied by (Shimodaira 2000; Sugiyama, Krauledat, and Müller 2007; Gretton et al. 2009). *Importance Weighting* is one of the prominently used techniques for tackling covariate shifts (Sugiyama, Krauledat, and Müller 2007; Lam, Li, and Prusty 2019). However, they are known to have high variance under minor shift scenarios (Cortes, Mansour, and Mohri 2010). Recently methods that emerged as the de-facto approaches to tackle distribution shifts include popular entropy minimization (Wang et al. 2021), pseudo-labeling (French, Mackiewicz, and Fisher 2017; Xie et al. 2020), batch normalization adaptation (Schneider et al. 2020; Nado et al. 2020), because of their wide applicability and superior performance. Our work provides a connection between a version of weighted entropy minimization and traditional importance sampling based loss which may be of independent interest.

Fairness under Distribution shift: The work by (Rezaei et al. 2021) is by far the most aligned to ours as they propose a method that is robust to covariate shift while ensuring fairness when unlabeled test data is available. However, this requires the density estimation of training and test distribution that is not efficient at higher dimensions and small number of test samples. In contrast our method avoids density estimation and uses a weighted version of entropy minimization that is constrained suitably to reflect importance sampling ratios implicitly. (Mandal et al. 2020) proposed a method for fair classification under the worst-case weighting of the data via an iterative procedure, but it is in the agnostic setting where test data is not available. (Singh et al. 2021) studied fairness under shifts through a causal lens but the method requires access to the causal graph, separating sets and other non-trivial data priors. (Zhang et al. 2021) proposed FARF, an adaptive method for learning in an online setting under fairness constraints, but is clearly different from the static shift setting considered in our work. (Slack, Friedler, and Givental 2020) proposed a MAML based algorithm to learn under fairness constraints, but it requires access to labeled test data. (An et al. 2022) propose a consistency regularization technique to ensure fairness under subpopulation and domain shifts under a specific model, while we consider covariate shift.

3 Problem Setup

Let $\mathcal{X} \subseteq \mathcal{R}^d$ be the d dimensional feature space for covariates, $\mathcal{A}$ be the space of categorical *group* attributes and $\mathcal{Y}$ be the space of class labels. In this work, we consider $\mathcal{A} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1\}$. Let $X \in \mathcal{X}, \mathrm{A} \in \mathcal{A}$, $\mathrm{Y} \in \mathcal{Y}$ be realizations from the space. We consider a training dataset $\mathcal{D}^S = \{(X_i, \mathrm{A}_i, \mathrm{Y}_i) | i \in [n]\}$ where every tuple $(X_i, \mathrm{A}_i, \mathrm{Y}_i) \in \mathcal{X} \times \mathcal{A} \times \mathcal{Y}$. We also have an *unlabeled* test dataset, $\mathcal{D}^T = \{X_i, \mathrm{A}_i | i \in [m]\}$. We focus on the setup where $m << n$. The training samples $(X_i, A_i, Y_i \in \mathcal{D}^S)$ are sampled i.i.d from distribution $\mathbb{P}^S(X, \mathrm{Y}, \mathrm{A})$ while the unlabeled test instances are sampled from $\mathbb{P}^T(X, \mathrm{A})$.

Let $\mathcal{F} : \mathcal{X} \to [0, 1]$ be the space of soft prediction models. In this work, we will consider $\mathsf{F} \in \mathcal{F}$ of the form $\mathsf{F} = h \circ g$ where $g(X) \in \mathbb{R}^k$ (for some dimension $k > 0$), is a representation that is being learnt while $h(g(X)) \in [0, 1]$ provides the soft prediction. Note that we don't consider A as an input to F, as explained in the work of (Zhao 2021). F is assumed to be parametrized via θ. Instead of representing the network as F_θ, we drop the subscript and simply use F when

Method	Metrics Opt.	Labels	Method Description
Adv-Deb (Zhang, Lemoine, and Mitchell 2018)	Eq. Odds	Yes	Rep. Matching (across subgroups grouped by (A,Y))
Adv-Deb (Zhang, Lemoine, and Mitchell 2018)	Dem. Parity	No	Rep. Matching (across subgroups grouped by only A)
FairFictPlay (Kearns et al. 2018)	FPR Parity	Yes	Minimax Game between learner and auditor
LFR (Swersky, Pitassi, and Dwork 2013)	Dem. Parity	No	Representation Matching
RSF* (Rezaei et al. 2021)	Eq. Odds	No	Minimax Game between predictor & test approximator
Massaging (Kamiran and Calders 2012)	Dem. Parity	Yes	Changing class labels to remove discrimination
Suppression (Kamiran and Calders 2012)	Dem. Parity	Yes	Remove sensitive & highly correlated attributes
Reweighing (Kamiran and Calders 2012)	Dem. Parity	Yes	Tuples in training dataset are assigned weights
Sampling (Kamiran and Calders 2012)	Dem. Parity	Yes	Non-uniform sampling via duplication & removal
RF* (Mandal et al. 2020)	Dem. Parity	No	Weighted Combination of Dataset via 2 Player Game
Opt. Pre-Processing (Calmon et al. 2017)	Dem. Parity	Yes	Data Transformation
Certify & Comb. Repair (Feldman et al. 2015)	Disparate Imp.	Yes	New distribution via linear interpolation in rank space
Certify & Geo. Repair (Feldman et al. 2015)	Disparate Imp.	Yes	New distribution via linear interpolation in original space
LAFTR (Madras et al. 2018)	Dem. Parity	No	Adversarial Representation Matching
LAFTR (Madras et al. 2018)	Eq. Odds	Yes	Adversarial Representation Matching
LAFTR (Madras et al. 2018)	Eq. Opp.	Yes	Adversarial Representation Matching
EGR (Agarwal et al. 2018)	Dem. Parity	No	Exp. Gradient Reduction
EGR (Agarwal et al. 2018)	Eq. Odds	Yes	Exp. Gradient Reduction
Post Processing (Hardt, Price, and Srebro 2016)	Eq. Odds	Yes	Modifying an existing predictor based on A and Y.

Table 1: Representative Collection of Fairness Methods. * represents methods that tackle covariate shift that we compare with

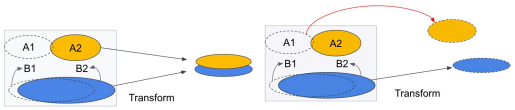

Figure 1: Asymmetric Shift Illustrated.

its clear from the context. The class prediction probabilities from F are denoted with $P(\hat{\mathrm{Y}} = y|X_i)$, where $y \in \{0, 1\}$.

The supervised in-distribution training of F is done by minimizing the *empirical risk*, $\widehat{\mathsf{ER}}^S$ as the proxy for *population risk*, $\mathcal{R}^S$. Both risk measures are computed using the *Cross Entropy (CE)* loss for classification (correspondingly we use $\widehat{\mathsf{ER}}^T$ and $\mathcal{R}^T$ over the *test distribution* for F).

$$\mathcal{R}^S = \mathbb{E}_{\mathbb{P}^S(X,\mathrm{A},Y)}\left(-\log P(\hat{\mathrm{Y}} = Y|X)\right),$$
$$\widehat{\mathsf{ER}}^S = \frac{1}{n}\sum_{(X_i,Y_i,A_i)\in\mathcal{D}^S}\left(-\log P(\hat{\mathrm{Y}} = Y_i|X_i)\right) \quad (1)$$

3.1 Covariate Shift Assumption

For our work, we adopt the *covariate shift* assumption as in (Shimodaira 2000). Covariate shift assumption implies that $\mathbb{P}^S(\mathrm{Y}|X, \mathrm{A}) = \mathbb{P}^T(\mathrm{Y}|X, \mathrm{A})$. In other words, shift in distribution only affects the joint distribution of covariates and sensitive attribute, i.e. $\mathbb{P}^S(X, \mathrm{A}) \neq \mathbb{P}^T(X, \mathrm{A})$. We note that our setup is identical to a recent work of fairness under covariate shift by (Rezaei et al. 2021). We also define and focus on a special case of covariate shift called *asymmetric covariate shift*.

Definition 3.1 (Asymmetric Covariate Shift)**.** Asymmetric covariate shift occurs when distribution of covariates of one group shifts while the other does not, i.e. $\mathbb{P}^T(X|\mathrm{A} = 1) \neq \mathbb{P}^S(X|\mathrm{A} = 1)$ while $\mathbb{P}^T(X|\mathrm{A} = 0) = \mathbb{P}^S(X|\mathrm{A} = 0)$ in addition to $\mathbb{P}^S(\mathrm{Y}|X, \mathrm{A}) = \mathbb{P}^T(\mathrm{Y}|X, \mathrm{A})$

This type of covariate shift occurs when a sub-group is over represented (sufficiently capturing all parts of the domain of interest in the training data) while the other sub-group being under represented and observed only in one part of the domain. In the test distribution, covariates of the under-represented group assume a more drastic shift.

Figure 1 shows how feature matching transformation affects domain of features for the two groups (yellow and the blue). The test features $B2, A2$ of both groups are made to overlap in the transformed space the training features. $A1, B1$ misalign in the transformed space of right subfigure when the same transformation as in the left subfigure is applied. Since $B1$ and $B2$ shift very little, in the transformed space, in the test region, only the labeled examples of blue group is found.

3.2 Fairness Regularization with no Labels

We observe that the central issue in our problem is the learner has access only to an unlabelled test set $\mathcal{D}^T$ and one wants some fairness criterion to be enforced on it. Popular fairness metrics like *Equalized Odds* (that forces $I(A; \hat{Y}|Y)$ to zero) and *Accuracy Parity* (that forces $I(Y \neq \hat{Y}; A) = 0$) depend on the training labels Y to even evaluate.

In Table 1, we outline fairness metrics optimized, methods used and popular methods in the fairness literature. We emphasise that this is by no means an exhaustive survey of fairness methods and only a representative one. Excluding methods that address accuracy-fairness tradeoff under covariate shift (we compare against these baselines empirically in our work), we observe that only methods that use representation matching of covariates across sensitive groups can work to impose any fairness measure at all without requiring any label information. We also point out that representation matching can also help ensure accuracy parity under some conditions (Zhao and Gordon 2019) without requiring labels. Therefore, we adopt representation matching loss across sensitive groups applied on $\mathcal{D}^T$ to be our fairness regularizer in

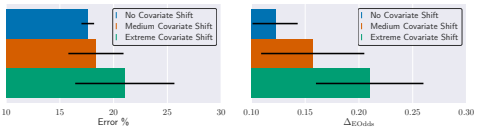

Figure 2: Both Error (in % left) & Eq. Odds (right) for SOTA Adv. Deb. exhibit strong degradation on increasing the magnitude of covariate shift. Three scenarios corresponding to no, intermediate and high shift are plotted, details in Sec 5.1

this work.

Formally, we seek to train a classifier $\mathsf{F}_\theta = h_\theta \circ g_\theta(X)$ by matching representation $g(X)$ across the protected sub groups and learning a classifier on top of that representation (Zhao and Gordon 2019). Several variants for representation matching loss have been proposed in the literature (Jiang et al. 2020; Wang, Nguyen, and Hanasusanto 2021; Zhao 2021; Chzhen et al. 2020). For implementation ease, we pick Wasserstein-2 metric to impose representation matching. We recall the definition of Wasserstein distance:

Definition 3.2. Let $(\mathcal{M}, d)$ be a metric space and $P_p(\mathcal{M})$ denote the collection of all probability measures μ on $\mathcal{M}$ with finite p^{th} moment. Then the p-th Wasserstein distance between measures μ and ν both $\in P_p(\mathcal{M})$ is given by: $\mathcal{W}_p(\mu,\nu) = \left(\inf_\gamma \int_{\mathcal{M}\times\mathcal{M}} d(x,y)^p d\gamma(x,y)\right)^{\frac{1}{p}}$; $\gamma \in \Gamma(\mu,\nu)$, where $\Gamma(\mu,\nu)$ denotes the collection of all measures on $\mathcal{M}\times\mathcal{M}$ with marginals μ and ν respectively.

We minimize the $\mathcal{W}_2$ between the representation $g(\cdot)$ of the test samples from both groups. Empirically, our representation matching loss is given by: $\hat{\mathcal{L}}_{Wass}(\mathcal{D}^T) = \mathcal{W}_p(\hat{\mu},\hat{\nu}), \hat{\mu} = \frac{\sum_{(\mathrm{X}_i, A_i=0)\in\mathcal{D}^T} \delta_{g(X_i)}}{|(\mathrm{X}_i, A_i=0)\in\mathcal{D}^T|}, \hat{\nu} = \frac{\sum_{(\mathrm{X}_i, A_i=1)\in\mathcal{D}^T} \delta_{g(X_i)}}{|(\mathrm{X}_i, A_i=1)\in\mathcal{D}^T|}$

We arrive at the following objective which is of central interest in the paper:

$$\min_{F_\theta = h_\theta \circ g_\theta} \widehat{\mathsf{ER}}^T + \lambda \hat{\mathcal{L}}_{Wass}(\mathcal{D}^T) \tag{2}$$

3.3 Issues with Applying Existing Techniques

Representation Matching: Since we don't have labels for the test set one cannot implement the first term in (2). It is natural to optimize $\widehat{\mathsf{ER}}^S + \lambda \hat{\mathcal{L}}_{Wass}(\mathcal{D}^T)$. Here, the first term optimizes prediction accuracy on labeled training data while the second term matches representation across groups in the unlabeled test. We illustrate that under asymmetric covariate shift, this above objective is ineffective. The issue is illustrated best through Figure 1. $A1$ and $B1$ represent group 1 and 0 feature distributions in the training set. Under asymmetric covariate shift, $B2 \approx B1$ while group 1 shifts drastically to $A2$. Now, representation matching loss on the test would map $A2$ and $B2$ to the same region in the range space of $g(\cdot)$ as in left subfigure of Figure 1. However, the classifier h would be exclusively trained on samples from group B (i.e. $g(B1)$) although both $A2$ and $B2$ overlap there as shown as in right subfigure of Figure 1. Training predictors on training samples but representation matching under test creates this issue (see (Havaldar et al. 2024)'s appendix). It also highlights the central issue of our paper. Adversarial debiasing (Zhang, Lemoine, and Mitchell 2018) is another method that matches representations, as discuss below.

Under-Performance of Adversarial Debiasing under Covariate Shift: We study the variation of the performance of a State of The Art Method, namely Adversarial Debiasing (Zhang, Lemoine, and Mitchell 2018) against the magnitude of shift γ on the Adult dataset. The variation of the error % is plotted in the left subfigure of Figure 2, and the variation of Equalized Odds is present in the right subfigure.

As we move from a setting of No Covariate Shift ($\gamma = 0$), to Medium Covariate Shift ($\gamma = 10$), to Extreme Covariate Shift ($\gamma = 20$), there is a significant deterioration in the performance of the method as both Error % and Equalized Odds increase with the increase in the magnitude of the shift. This reinforces the belief that state of the art fairness techniques do not extend well to settings under covariate shift. In figure 2, we complement these claims by analyzing the *under-performance* for a state-of-the-art fairness method - Adversarial Debiasing (Zhang, Lemoine, and Mitchell 2018). We also see similar drop in performance under covariate shift in other baselines we consider, which we have highlighted in our experimental analysis.

Distributional Robustness Methods: Another option to implement (2) would be to use a distributional robust learner (DRO) on the source distribution simultaneously with the representation matching penalty for the target. We consider a very recent SOTA method RGD-Exp (Kumar et al. 2023) that implements a form of DRO. We effectively replace $\widehat{\mathsf{ER}}^T$ from eqn (2) with a robust loss term from the paper and perform the same optimization as us and notice that it does not achieve as good a accuracy-fairness tradeoff as our algorithm, thus establishing that trivially combining a SOTA distributionally robust method with Wasserstein Loss (2nd term from eqn. 2: $\hat{\mathcal{L}}_{Wass}(\mathcal{D}^T)$) does not suffice to achieve fairness under shift and something more nuanced is required.

Importance Sampling/Density Ratio Estimation based methods: Another way to implement (2) is to use importance sampled prediction loss on training samples to mimic the test loss (first term) in (2). For this, one estimates ratio between training and test density directly using KLIEP/LSIF losses (Sugiyama, Krauledat, and Müller 2007; Kanamori, Hido, and Sugiyama 2009) or perform density estimation which does not scale in higher dimensions. Sample complexity of these techniques directly scales with importance sampling variance which is large with very few test samples. We show this via formal sample complexity bounds in Section 4.1 and empirically in figure 4 and others in (Havaldar et al. 2024)'s appendix, where we see large variances in accuracy for these methods. Robust Shift Fair from (Rezaei et al. 2021) also involves density estimation steps which suffer from the same.

4 Method and Algorithm

Recall that the objective we are interested in is (2). One needs a proxy for the first term ($\widehat{\mathsf{ER}}^T$) due to lack of labels. From considerations in the previous section, training has

to be done in a manner that exploits training labels from source dataset effectively but can tackle covariate shift despite using representation matching. We derive a novel objective in Theorem 4.1 based on the weighted entropy over instances in $\mathcal{D}^T$ along with empirical loss over $\mathcal{D}^S$ and show that is an upper bound to $\mathcal{R}^T$.

Theorem 4.1. *Suppose that $\mathbb{P}^T(\cdot)$ and $\mathbb{P}^S(\cdot)$ are absolutely continuous with respect to each other over domain $\mathcal{X}$. Let $\epsilon \in \mathbb{R}^+$ be such that $\frac{\mathbb{P}^T(\mathrm{Y}=y|X)}{P(\hat{\mathrm{Y}}=y|X)} \leq \epsilon$, for $y \in \{0,1\}$ almost surely with respect to distribution $\mathbb{P}^T(X)$. Then, we can upper bound $\mathcal{R}^T$ using $\mathcal{R}^S$ along with an unsupervised objective over $\mathbb{P}^T$ as:*

$$\mathcal{R}^T \leq \mathcal{R}^S + \epsilon \times \mathbb{E}_{\mathbb{P}^T(X)}\left[e^{\left(-\frac{\mathbb{P}^S(X)}{\mathbb{P}^T(X)}\right)}\mathcal{H}(\hat{\mathrm{Y}}|\mathcal{X})\right] \tag{3}$$

where, $\mathcal{H}(\hat{\mathrm{Y}}|\mathcal{X}) = \sum_{y\in\{0,1\}} -P(\hat{\mathrm{Y}}=y|X)\log(P(\hat{\mathrm{Y}}=y|X))$ is conditional entropy of the label given a sample X.

Proof. Refer (Havaldar et al. 2024)'s appendix for proof. □

Note: Sections marked as A.x, B.y & C.z refer to (Havaldar et al. 2024)'s supplement. The mild assumption on ϵ in the theorem is also justified via extensive experiments in B.4.3

We *emphasize* that this result also provides an important connection and a rationale for using entropy based objectives as an unsupervised adaptation objective from an importance sampling point of view that has been missing in the literature (Wang et al. 2021; Sun et al. 2019).

Entropy objective is imposed on points that are more typical with respect to the test than the training. Conversely, in the region where samples are less likely with respect to the test distribution, since it has been optimized for label prediction as part of training, the entropy objective is not imposed strongly. The above bound however hinges on the assumption that point-wise in the domain $\mathcal{X}$, F approximates the true soft predictor by at most a constant factor ϵ. To ensure a small value of ϵ, we resort to pre-training F with only $\mathcal{D}^S$ samples for a few epochs before imposing any other type of regularization.

4.1 Theoretical Analysis

The most widely used objective to optimize for (Left Hand Side) L.H.S of (3), i.e. $\mathcal{R}^T$, leverages *importance sampling* (Sugiyama, Krauledat, and Müller 2007), which we denote as $\mathcal{R}_{IS}$ here for clarity. We denote R.H.S of (3) by $\mathcal{R}_{WE}$. Our method is motivated by the R.H.S of (3). Here, we compare the generalization bounds for $\mathcal{R}_{IS}$ and $\mathcal{R}_{WE}$. We make the following assumptions to simplify the analysis as the task is to compare $\mathcal{R}_{IS}$ against $\mathcal{R}_{WE}$ only, however some of these can be relaxed trivially.

Assumption 4.2.
- Let $\Theta = \{\theta_1 \dots \theta_k\}$ be finite parameter space.
- Let the losses $l_1(\cdot) = -log(P_\theta(\hat{Y}=Y|X))$ and $l_2(\cdot) = \sum_{y\in\{0,1\}} -P_\theta(\hat{Y}=y|X)\log P_\theta(\hat{Y}=y|X)$ be bounded between $[0,1]$ in the domain $\{0,1\}\times\mathcal{X}$ for all $\theta\in\Theta$. This is not a heavy assumption and can be achieved via appropriate Lipschitz log loss over bounded domain.
- Denoting the important weights $z(X) = \frac{\mathbb{P}^T(X)}{\mathbb{P}^S(X)}$ (assuming we have access to exact importance weights), let $\sup_{X\in\mathcal{X}} z(X) = M$ and the variance of the weights with respect to the training distribution be σ^2 .

For the $\mathcal{R}_{IS}$ objective, we have the following result,

Theorem 4.3. *Under Assumption 4.2, with probability $1-\delta$ over the draws of $\mathcal{D}^S \sim \mathbb{P}^S$, we have $\forall\theta\in\Theta$: $\mathbb{E}_{\mathbb{P}^S}[\mathcal{R}_{IS}(\theta)] \leq \widehat{\mathcal{R}}_{IS}(\theta) + \frac{2M(\log|\Theta|+\log(1/\delta))}{3|\mathcal{D}^S|} + \sqrt{2\sigma^2\frac{(\log|\Theta|+\log(1/\delta))}{|\mathcal{D}^S|}}$*

Whereas for our obj. $\mathcal{R}_{WE}$ (posing ϵ as a hyperparam λ),

Theorem 4.4. *Under Assumption 4.2, we have that with probability $1-2\delta$ over the draws of $\mathcal{D}^S \sim \mathbb{P}^S$ and $\mathcal{D}^T \sim \mathbb{P}^T$, we have $\forall\theta\in\Theta$ $\mathbb{E}_{\mathbb{P}^S,\mathbb{P}^T}[\mathcal{R}_{WE}(\theta)] \leq \widehat{\mathcal{R}}_{WE}(\theta) + 2\sqrt{\frac{2\log|\Theta|}{|\mathcal{D}^S|}} + 2\lambda\sqrt{\frac{2\log|\Theta|}{|\mathcal{D}^T|}} + 3\sqrt{\frac{\ln(2/\delta)}{2|\mathcal{D}^S|}} + 3\lambda\sqrt{\frac{\ln(2/\delta)}{2|\mathcal{D}^T|}}$*

Refer (Havaldar et al. 2024)'s B.5 for proofs. Comparing Theorem 4.3 to Theorem 4.4, we see that the generalization bound for importance sampled objective, $\mathcal{R}_{IS}$, depends on variance of importance weights σ^2 and the worst case value M. In contrast, our objective, $\mathcal{R}_{WE}$, *does not* depend on these parameters and thus does not suffer from high variance. These results are further justified empirically in section 5.2.

4.2 Weighted Entropy Objective

Implementing the objective in (3), requires computation of $\frac{\mathbb{P}^S(X)}{\mathbb{P}^T(X)}$. This is challenging when m (amount of unlabeled test samples) is small and typical way of density estimation in high dimensions is particularly hard. Therefore, we propose to estimate the ratio $\frac{\mathbb{P}^S(X)}{\mathbb{P}^T(X)}$ by a parametrized network $\mathsf{F}_w : \mathcal{X}\to\mathbb{R}$, where $\mathsf{F}_w(X)$ shall satisfy the following constraints: $\mathbb{E}_{X\sim\mathbb{P}^T(X)}[\mathsf{F}_w(X)] = 1$, and $\mathbb{E}_{X\sim\mathbb{P}^S(X)}[1/(\mathsf{F}_w(X))] = 1$. By definition, these constraints must be satisfied.

Building on (3), we solve for the following upper bound in Theorem 4.1:

$$\max_{\theta(\mathsf{F}_w)} \mathcal{R}^S + \epsilon \times \mathbb{E}_{\mathbb{P}^T(X)}\left[e^{(-\mathsf{F}_w(X))}\mathcal{H}(\hat{\mathrm{Y}}|\mathcal{X})\right] \text{ s.t.}$$
$$\mathbb{E}_{X\sim\mathbb{P}^T(X)}[\mathsf{F}_w(X)] = 1,\ \mathbb{E}_{X\sim\mathbb{P}^S(X)}[1/(\mathsf{F}_w(X))] = 1 \tag{4}$$

Final Learning Objective: Finally, we plug in the empirical risk estimator for $\mathcal{R}^S$, approximate the expectation in second term with the empirical version over $\mathcal{D}^T$, posit ϵ as a hyperparameter and add the unfairness objective $\hat{\mathcal{L}}_{Wass}(\mathcal{D}^T)$ as in (2). Furthermore, we utilize the output of the representation layer g (denoting $F = h\circ g$, where g is the encoder subnetwork and h is the classifier subnetwork, as input to F_w rather than raw input X (provable benefits of representation learning (Arora and Risteski 2017)). Hence, the optimization objective becomes:

$$\min_{\mathsf{F}_\theta}\max_{\mathsf{F}_w} \mathcal{L}(\mathsf{F}_\theta,\mathsf{F}_w) = \widehat{\mathsf{ER}}^S + \lambda_1\frac{1}{m}\sum_{X_i\in\mathcal{D}^T}[e^{(-\mathsf{F}_w(g(X_i)))}$$
$$\mathcal{H}(\hat{\mathrm{Y}}|\mathcal{X})] + \lambda_2\hat{\mathcal{L}}_{Wass}(\mathcal{D}^T) \text{s.t. } \mathcal{C}_1 = \frac{1}{m}\sum_{X_i\in\mathcal{D}^T}\mathsf{F}_w(g(X_i))$$
$$= 1,\ \text{and } \mathcal{C}_2 = \frac{1}{n}\sum_{X_i\in\mathcal{D}^S}\frac{1}{\mathsf{F}_w(g(X_i))} = 1 \tag{5}$$

Algorithm 1: Gradient Updates for the proposed objective to learn fairly under covariate shift

Input: Training data $\mathcal{D}^S$, Unlabelled Test data $\mathcal{D}^T$, model F, weight estimator F_w, decaying learning rate η_t, number of pre-training steps $\tilde{\mathcal{E}}$, number of training steps $\mathcal{E}$ for eq 5, λ_1, λ_2
Output: Optimized parameters θ^* of the model F

$\theta^0 \leftarrow$ random initialization
for $t \leftarrow 1$ **to** $\tilde{\mathcal{E}}$ **do**
  $\theta^t \leftarrow \theta^{t-1} - \eta_t \nabla_{\theta^{t-1}} \widehat{\mathsf{ER}}^S$
end for
$w^{\tilde{\mathcal{E}}} \leftarrow$ random initialization
for $t \leftarrow \tilde{\mathcal{E}}+1$ **to** $\mathcal{E}+\tilde{\mathcal{E}}$ **do**
  $w^t \leftarrow w^{t-1} + \eta_t \nabla_{w^{t-1}} \mathcal{L}(\theta^{t-1}, w^{t-1})$; /* subject to $\mathcal{C}_1$ and $\mathcal{C}_2$ */
  $\theta^t \leftarrow \theta^{t-1} - \eta_t \nabla_\theta \mathcal{L}(\theta^{t-1}, w^t)$; /* gradient stopping is applied through F_w in this step*/
end for
$\theta^* \leftarrow \theta^{\mathcal{E}+\tilde{\mathcal{E}}}$

Here λ_1 and λ_2 are hyperparameters governing the objectives. $\mathcal{C}_1$ and $\mathcal{C}_2$ are the constraints. We use alternating gradient updates to solve the above min-max problem. Our entire learning procedure consists of *two stages*: (1) pre-training F for some epochs with only $\mathcal{D}^S$ and (2) further training F with (5). The procedure is summarized in Algorithm 1 and a high level architecture is provided in (Havaldar et al. 2024)'s A.4

5 Experiments

We demonstrate our method on 4 widely used benchmarks in the fairness literature, i.e. Adult, Communities and Crime, Arrhythmia and Drug Datasets with detailed description in (Havaldar et al. 2024)'s appendix section A.2 due to space constraints. The baseline methods used for comparison are: MLP, Adversarial Debias (AD) (Zhang, Lemoine, and Mitchell 2018), Robust Fair (RF) (Mandal et al. 2020), Robust Shift Fair (RSF) (Rezaei et al. 2021), Z-Score Adaptation (ZSA). Along these, we also compare against two popular Density ratio estimation techniques, (Sugiyama et al. 2007) (KLIEP) and (Kanamori, Hido, and Sugiyama 2009) (LSIF), that estimate the ratio $\frac{\mathbb{P}^T(X)}{\mathbb{P}^S(X)}$ via a parametrized setup. The estimates are then used to compute the *importance weighted* training loss $\mathcal{R}_{IS}$ described previously. (Menon and Ong 2016) analysed both these methods in a unifying framework. We further provide comparisons against an adapted RGD-Exp (Kumar et al. 2023) as described in section 3.3 to demonstrate simply adapting DRO methods would not suffice. The detailed description for all the baselines is provided in (Havaldar et al. 2024)'s appendix section A.3. These baselines also cover the important works highlighted in Section 2.

The implementation details of all the methods with relevant hyperparameters are provided in (Havaldar et al. 2024)'s A.4 The evaluation of our method against the baselines is done via the trade-off between fairness violation (using Δ_{EOdds}) and error (which is $100-$ accuracy). All algorithms are run 50 times before reporting the mean and the standard deviation. All experiments are run on single NVIDIA Tesla V100 GPU.

Apart from the primary results on standard and asymmetric shift below, extensive analyses across multiple settings are provided in (Havaldar et al. 2024)'s appendix (due to space).

5.1 Shift Construction

To construct the covariate shift in the datasets, i.e., to introduce $\mathbb{P}^S(X, \mathrm{A}) \neq \mathbb{P}^T(X, \mathrm{A})$, we utilize the following strategy akin to the works of (Rezaei et al. 2021; Gretton et al. 2008). First, all the non-categorical features are normalized by *z-score*. We then obtain the *first principal component* of the of the covariates and further project the data onto it, denoting it by $\mathcal{P}_{\mathcal{C}}$. We assign a score to each point $\mathcal{P}_{\mathcal{C}}[i]$ using the density function $\Xi : \mathcal{P}_{\mathcal{C}}[i] \rightarrow \mathrm{e}^{\gamma \cdot (\mathcal{P}_{\mathcal{C}}[i] - b)}/\mathcal{Z}$. Here, γ is a hyperparameter controlling the level of distribution shift under the split, b is the $60^{th}‰$ (percentile) of $\mathcal{P}_{\mathcal{C}}$ and $\mathcal{Z}$ is the normalizing coefficient computed empirically. Using this, we sample 40% instances from the dataset as the test and remaining 60% as training. To construct the validation set, we further split the training subset to make the final train:validation:test ratio as $5:1:4$, where the test is distribution shifted. Similar procedure is used to construct the shifts for asymmetric analysis in section 5.3.

Note that for large values of γ, all the points with $\mathcal{P}_{\mathcal{C}[i]} > b$ will have high density thereby increasing the probability of being sampled into the test set. This generates a sufficiently large distribution shift. Correspondingly, for smaller values of γ, the probability of being sampled is not sufficiently high for these points thereby leading to higher overlap between the train and test distributions.

5.2 Fairness-Accuracy Tradeoff

The experimental results for the shift constructed using procedure in section 5.1 are shown in Figure 3. The results closer to the *bottom left* corner in each plot are desirable.

Our method provides a better error and fairness tradeoffs against the baselines on all benchmarks. For example, on the Adult dataset, we have the lowest error rate at around 15% with Δ_{EOdds} at almost 0.075 while the closest baselines MLP and RF fall short on either of the metrics. On Arrhythmia and Communities, our method achieves very low Δ_{EOdds} (best on Arrhythmia with a margin of $\sim 30\%$) with only marginally higher error as compared to MLP and RF respectively. On the Drug dataset, we achieve the best numbers for both the metrics. For the same accuracy, we obtain 1.3x-2x improvements against the baselines methods on most of the benchmarks. Similarly, for the same Δ_{EOdds}, we achieve up to 1.5x lower errors. It is also important to note that all the other unsupervised adaptation algorithms perform substantially worse and are highly unreliable. For example, ZSA performs well only on the Drug dataset, but shows extremely worse errors (even worse than *random predictions*) on Communities and Adult. The adaptation performed by ZSA is insufficient to handle covariate shift. RSF is consistently worse across the board. This is because it tries to explicitly estimate $\mathbb{P}^S(X)$ and $\mathbb{P}^T(X)$ which is extremely challenging whereas we implicitly estimate the importance ratio.

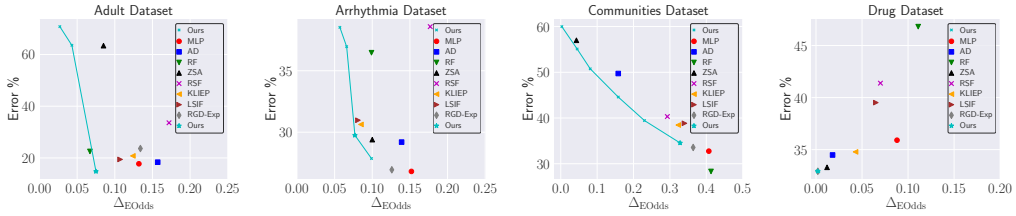

Figure 3: Fairness-Error Tradeoff Curves for our method (Pareto Frontier) against the optimal performance of the baselines. Our method provides better tradeoffs in all cases. (On Drug dataset, the performance is concentrated around the optimal point)

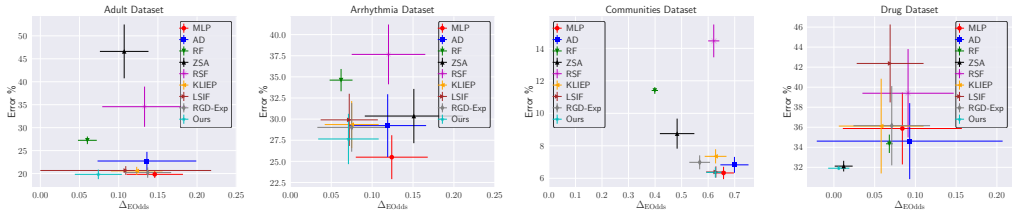

Figure 4: Comparison of our method against the baselines under Asymmetric Covariate Shift for group $A = 0$.

For KLIEP and LSIF, we equip both of these with the Wasserstein penalty term to provide a fair comparison. First, we observe that our method consistently outperforms these algorithms across the datasets. The relative improvement of our method is as high as $\sim 31\%$ in error on Adult dataset and $\sim 32.5\times$ in Δ_{EOdds} on Drug dataset against LSIF. Similar non-trivial margins can be noted on other datasets. Second, as highlighted in (Havaldar et al. 2024)'s appendix, the variance in error rates of the KLIEP and LSIF based importance is very high on the Drug dataset. Particularly, both KLIEP and LSIF exhibit up to $20 - 40$ times higher variance in error and up to $10 - 12$ times in Δ_{EOdds}. We can attribute this difference to the phenomenon that in the small sample regime, importance weighted objective on training dataset alone may not bring any improvements for covariate shift due to variance issues and thus estimating the ratio can be insufficient.

Variance Details: The detailed plots with variance corresponding to figure 3 are provided in appendix. In some cases, the standard deviation bars in the figure stretch beyond 0 in $\mathbb{R}^-$ due to skewness when error bars are plotted, however numbers across all the runs are *positive*. Low variance results of our method are notable, as discussed in section 5.2 especially against KLEIP and LSIF.

5.3 Results on Asymmetric Shift

Here, we present empirical results for Asymmetric Covariate Shift where the degree of shift is substantially different across the groups. To construct data for this setting, we follow the same procedure as described in section 5.1, but operate on data for the two groups differently. The shift is introduced in one of the groups while for the other group, we resort to splitting it randomly into train-val-test.

Figure 4 provides the results for the setup when shift is created in group $A = 0$. We again observe that our method provides better tradeoffs across the board. For the shift in group $A = 0$, we have substantially better results on Adult and Arrhythmia with up to ~ 2x improvements on Δ_{EOdds} for similar error and up to ~ 1.4x improvements in error for similar Δ_{EOdds}. On the Communities dataset, MLP and AD show similar performance to ours, but much worse on the Drug dataset for both the metrics. ZSA performs comparably to our method only on Drug, but is substantially worse on other datasets. This confirms the inconsistency of the baselines under this setup as well. The results for shift in group $A = 1$ (see (Havaldar et al. 2024)'s appendix), show analogous trends. On the Drug dataset we are **10x** and **5x** better than the two importance sampling baselines on Δ_{EOdds} without the significant variance and lower error % as well. Even on other datasets we notice strong trends for our method with lower error and lower Δ_{EOdds} across the board. This shows that our method performs well in the Asymmetric Covariate Shift setting against importance sampling methods. It is also important to note that the errors are lower for all the methods as compared to figure 3 since only one group exhibits substantial shift while degradation in equalized odds is higher. This is in line with the reasoning provided in section 3.2 based on theorem C.1 in (Havaldar et al. 2024)'s appendix.

6 Conclusion

In this work, we considered the problem of unsupervised test adaptation under covariate shift to achieve good fairness-error trade-offs using a small amount of unlabeled test data. We proposed a composite loss, that apart from prediction loss on training, involves a representation matching loss along with weighted entropy loss on the unsupervised test. We experimentally demonstrated the efficacy of our formulation. Our source code is made available for additional reference.[1]

[1] https://github.com/google/uafcs

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*, 60–69. PMLR.

An, B.; Che, Z.; Ding, M.; and Huang, F. 2022. Transferring Fairness under Distribution Shifts via Fair Consistency Regularization.

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Arora, S.; and Risteski, A. 2017. Provable benefits of representation learning.

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2006. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19.

Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Wortman, J. 2007. Learning bounds for domain adaptation. *Advances in neural information processing systems*, 20.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.

Calmon, F.; Wei, D.; Vinzamuri, B.; Natesan Ramamurthy, K.; and Varshney, K. R. 2017. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30.

Celis, L. E.; Huang, L.; Keswani, V.; and Vishnoi, N. K. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, 319–328.

Chzhen, E.; Denis, C.; Hebiri, M.; Oneto, L.; and Pontil, M. 2020. Fair regression with Wasserstein barycenters. In *Advances in Neural Information Processing Systems*.

Cortes, C.; Mansour, Y.; and Mohri, M. 2010. Learning Bounds for Importance Weighting. In Lafferty, J.; Williams, C.; Shawe-Taylor, J.; Zemel, R.; and Culotta, A., eds., *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

Cotter, A.; Jiang, H.; Gupta, M.; Wang, S.; Narayan, T.; You, S.; and Sridharan, K. 2019. Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals. *Journal of Machine Learning Research*, 20(172): 1–59.

Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J. S.; and Pontil, M. 2018. Empirical risk minimization under fairness constraints. *Advances in Neural Information Processing Systems*, 31.

Duchi, J. C.; and Namkoong, H. 2021. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3): 1378–1406.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268.

Fish, B.; Kun, J.; and Lelkes, Á. D. 2016. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM international conference on data mining*, 144–152. SIAM.

French, G.; Mackiewicz, M.; and Fisher, M. 2017. Self-ensembling for visual domain adaptation. *arXiv*.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 1180–1189. PMLR.

Gretton, A.; Smola, A.; Huang, J.; Schmittfull, M.; Borgwardt, K.; and Schölkopf, B. 2009. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4): 5.

Gretton, A.; Smola, A.; Huang, J.; Schmittfull, M.; Borgwardt, K.; and Schölkopf, B. 2008. Covariate Shift by Kernel Mean Matching. In *Dataset Shift in Machine Learning*. The MIT Press. ISBN 9780262170055.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Havaldar, S.; Chauhan, J.; Shanmugam, K.; Nandy, J.; and Raghuveer, A. 2024. Fairness under Covariate Shift: Improving Fairness-Accuracy tradeoff with few Unlabeled Test Samples. arXiv:2310.07535.

Jiang, R.; Pacchiano, A.; Stepleton, T.; Jiang, H.; and Chiappa, S. 2020. Wasserstein Fair Classification. In *UAI*.

Kamiran, F.; and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1): 1–33.

Kanamori, T.; Hido, S.; and Sugiyama, M. 2009. A Least-squares Approach to Direct Importance Estimation. *Journal of Machine Learning Research*, 10(48): 1391–1445.

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, 2564–2572. PMLR.

Kumar, R.; Majmundar, K.; Nagaraj, D.; and Suggala, A. S. 2023. Stochastic Re-weighted Gradient Descent via Distributionally Robust Optimization. *arXiv preprint arXiv:2306.09222*.

Lam, H.; Li, F.; and Prusty, S. 2019. Robust Importance Weighting for Covariate Shift.

Li, C.; Lu, Y.; Mei, Q.; Wang, D.; and Pandey, S. 2015. Click-through prediction for advertising in twitter timeline. In *KDD*.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, 3384–3393. PMLR.

Mandal, D.; Deng, S.; Jana, S.; Wing, J.; and Hsu, D. J. 2020. Ensuring Fairness Beyond the Training Data. In *Advances in Neural Information Processing Systems*, volume 33.

Menon, A.; and Ong, C. S. 2016. Linking losses for density ratio and class-probability estimation. In *Proceedings of The 33rd International Conference on Machine Learning*.

Mitchell, S.; Potash, E.; Barocas, S.; D'Amour, A.; and Lum, K. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8: 141–163.

Moreno-Torres, J. G.; Raeder, T.; Alaiz-Rodríguez, R.; Chawla, N. V.; and Herrera, F. 2012. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1): 521–530.

Nado, Z.; Padhy, S.; Sculley, D.; D'Amour, A.; Lakshminarayanan, B.; and Snoek, J. 2020. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*.

Oneto, L.; and Chiappa, S. 2020. Fairness in machine learning. In *Recent Trends in Learning From Data*, 155–196. Springer.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On fairness and calibration. *Advances in neural information processing systems*, 30.

Redko, I.; Morvant, E.; Habrard, A.; Sebban, M.; and Bennani, Y. 2020. A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv preprint arXiv:2004.11829*.

Rezaei, A.; Liu, A.; Memarrast, O.; and Ziebart, B. D. 2021. Robust fairness under covariate shift. In *AAAI*.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Schneider, S.; Rusak, E.; Eck, L.; Bringmann, O.; Brendel, W.; and Bethge, M. 2020. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems*.

Shi, Y.; Seely, J.; Torr, P. H.; Siddharth, N.; Hannun, A.; Usunier, N.; and Synnaeve, G. 2021. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*.

Shimodaira, H. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2): 227–244.

Silva, S. H.; and Najafirad, P. 2020. Opportunities and challenges in deep learning adversarial robustness: A survey. *arXiv preprint arXiv:2007.00753*.

Singh, H.; Singh, R.; Mhasawade, V.; and Chunara, R. 2021. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 3–13.

Slack, D.; Friedler, S. A.; and Givental, E. 2020. Fairness warnings and Fair-MAML: learning fairly with minimal data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 200–209.

Sugiyama, M.; Krauledat, M.; and Müller, K.-R. 2007. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5).

Sugiyama, M.; Nakajima, S.; Kashima, H.; Buenau, P.; and Kawanabe, M. 2007. Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation. In *Advances in Neural Information Processing Systems*, volume 20.

Sun, Y.; Wang, X.; Liu, Z.; Miller, J.; Efros, A. A.; and Hardt, M. 2019. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts.

Swersky, R. Z. Y. W. K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. *ICML. PMLR*.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *ICLR*.

Wang, Y.; Nguyen, V. A.; and Hanasusanto, G. A. 2021. Wasserstein Robust Classification with Fairness Constraints.

Wiles, O.; Gowal, S.; Stimberg, F.; Alvise-Rebuffi, S.; Ktena, I.; Cemgil, T.; et al. 2021. A fine-grained analysis on distribution shift. *arXiv preprint arXiv:2110.11328*.

Wilson, G.; and Cook, D. J. 2020. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5): 1–46.

Xie, Q.; Luong, M.-T.; Hovy, E.; and Le, Q. V. 2020. Self-training with noisy student improves imagenet classification. In *CVPR*.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *In WWW*.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.

Zhang, W.; Bifet, A.; Zhang, X.; Weiss, J. C.; and Nejdl, W. 2021. Farf: A fair and adaptive random forests classifier. In *KDD*.

Zhao, H. 2021. Costs and Benefits of Fair Regression.

Zhao, H.; and Gordon, G. 2019. Inherent tradeoffs in learning fair representations. *Advances in neural information processing systems*, 32.