Robust Fairness under Covariate Shift

Ashkan Rezaei[1], **Anqui Liu**[2], **Omid Memarrast**[1], **Brian Ziebart**[1]
[1] Department of Computer Science, University of Illinois at Chicago
[2] California Institute of Technology
arezae4@uic.edu, anqiliu@caltech.edu, omemar2@uic.edu, bziebart@uic.edu

Abstract

Making predictions that are fair with regard to protected group membership (race, gender, age, etc.) has become an important requirement for classification algorithms. Existing techniques derive a fair model from sampled labeled data relying on the assumption that training and testing data are identically and independently drawn (iid) from the same distribution.

In practice, distribution shift can and does occur between training and testing datasets as the characteristics of individuals interacting with the machine learning system—and which individuals interact with the system—change. We investigate fairness under covariate shift, a relaxation of the iid assumption in which the inputs or covariates change while the conditional label distribution remains the same. We seek fair decisions under these assumptions on target data with unknown labels. We propose an approach that obtains the predictor that is robust to the worst-case in terms of target performance while satisfying target fairness requirements and matching statistical properties of the source data. We demonstrate the benefits of our approach on benchmark prediction tasks.

Introduction

Supervised learning algorithms typically focus on optimizing one singular objective: predictive performance on unseen data. However, the social impact of unwanted bias in these algorithms has become increasingly important. Machine learning systems that disadvantage specific groups are less likely to be accepted and may violate disparate impact law (Chang 2006; Kabakchieva 2013; Lohr 2013; Shipp et al. 2002; Obermeyer and Emanuel 2016; Moses and Chan 2014; Shaw and Gentry 1988; Carter and Catlett 1987; O'Neil 2016). Fairness through unawareness, which simply denies knowledge of protected group membership to the predictor, is insufficient to effectively guarantee fairness because other characteristics or covariates may correlate with protected group membership (Pedreshi, Ruggieri, and Turini 2008). Thus, there has been a surge of interest in the machine learning community to define fairness requirements reflecting desired behavior and to construct learning algorithms that more effectively seek to satisfy those requirements in various settings (Mehrabi et al. 2019; Barocas, Hardt, and Narayanan 2017; Calmon et al. 2017; Donini et al. 2018; Dwork et al. 2012, 2017; Hardt, Price, and Srebro 2016; Zafar et al. 2017a; Zemel et al. 2013; Jabbari et al. 2016; Chierichetti et al. 2017).

Though many definitions and measures of (un)fairness have been proposed (See Verma and Rubin (2018); Mehrabi et al. (2019)), the most widely adopted are group fairness measures of demographic parity (Calders, Kamiran, and Pechenizkiy 2009), equalized opportunity, and equalized odds (Hardt, Price, and Srebro 2016).

A number of techniques have been developed as either post-processing steps (Hardt, Price, and Srebro 2016) or in-processing learning methods (Agarwal et al. 2018; Zafar et al. 2017a; Rezaei et al. 2020) seeking to achieve fairness according to these group fairness definitions. These methods attempt to make fair predictions at testing time by relying heavily on an assumption that training and testing data are *independently and identically drawn (iid)* from the same distribution, so that providing fairness on the training dataset provides approximate fairness on the testing dataset.

In practice, it is common for data distributions to *shift* between the training data set (*source distribution*) and the testing data set (*target distribution*). For example, the characteristics of loan applicants may differ significantly over time due to macroeconomic trends or changes in the self-selection criteria that potential applicants employ. Fairness methods that ignore such shifts may satisfy definitions for fairness on training samples, while violating those definitions severely on testing data. Indeed, disparate performance for underrepresented groups in computer vision tasks has been attributed to manually labeled data that is highly biased compared to testing data (Yang et al. 2020). Explicitly incorporating these shifts into the design of predictors is crucial for realizing fairer applications of machine learning in practice. However, the resulting problem setting is particularly challenging; access to labels is only available for the training distribution. Fair prediction methods could fail by using only source labels, especially for fairness definitions that condition on ground-truth labels, like equal opportunity. Figure 1 illustrates the declining performance of in-processing method of (Rezaei et al. 2020) that do not incorporate any consideration of distribution shift and instead only depend on source fairness measurements. As the amount of shift increases (x axis), the fairness of the predictor (y axis) decreases. Therefore, relying on the iid assumption, which is often violated in practice, introduces significant limitations for realizing desired fairness in critical applications

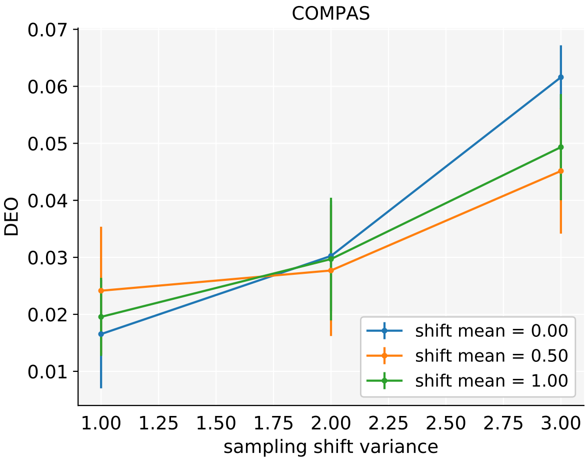


Figure 1: The fairness constraint violation (differnce between two groups) on the target distribution of fairLR (Rezaei et al. 2020), that do not account for distribution shift, and corrects for equalized opportunity using source data. The violation increases as the shift intensity increases. Experiment section includes details about simulating the shift.

We seek to address the task of providing fairness guarantees under the non-iid assumption of covariate shift. Covariate shift is a special case of data distribution shift. It assumes that the relationship between labels and covariates (inputs) is the same for both distributions, while only the source and target covariate distributions differ. Under the fair prediction setting, the sensitive group features are usually correlated with other features. Covariate shift then indicates that the labels given the covariates, including the sensitive group features, stays the same between two distributions. For example, even though there are less female loan applicants in area A than area B, which causes a marginal input distribution to shift between these two areas, we believe the probability of belonging to the advantages class (e.g., repaying a loan) given the full covariate should be the same.

In this paper, we propose a robust estimation approach for constructing a fair predictor under covariate shift. We summarize our contribution as follows:

We formulate the fair prediction problem as a game between an adversary choosing conditional label distributions to fairly minimize predictive loss on the target distribution, and an adversary choosing conditional label distributions that maximize that same objective. Constraints on the adversary require it to match statistics under the source distribution. Fairness is incorporated into the formulation by a penalty term in the objective that evaluates the fairness on the target input distribution and adversary's conditional label distribution.

We derive a convex optimization problem from the formulation and obtain the predictor's and adversary's conditional label distribution that are parametric, but cannot be solved analytically. Based on the formulation, we propose a batch gradient descent algorithm for learning the parameters.

We compare our proposed method with baselines that only account for covariate shift or fairness or both. We demonstrate that our method outperforms the baselines on both predictive performance and the fairness constraints satisfaction under covariate shift settings.

Related Work

Fairness

Various methods have been developed in recent years to achieve fair classification according to group fairness definitions. These techniques can be broadly categorized as pre-processing modifications to the input data or fair representation learning (Kamiran and Calders 2012; Calmon et al. 2017; Zemel et al. 2013; Feldman et al. 2015; Del Barrio et al. 2018; Donini et al. 2018; Guidotti et al. 2018), post-processing correction of classifiers' outputs (Hardt, Price, and Srebro 2016; Pleiss et al. 2017), in-processing methods that incorporate the fairness constraints into the training process and parameter learning procedure (Donini et al. 2018; Zafar et al. 2017c,a,b; Cotter et al. 2018; Goel, Yaghini, and Faltings 2018; Woodworth et al. 2017; Kamishima, Akaho, and Sakuma 2011; Bechavod and Ligett 2017; Quadrianto and Sharmanska 2017; Rezaei et al. 2020), meta-algorithms (Celis et al. 2019; Menon and Williamson 2018), reduction-based methods (Agarwal et al. 2018; Cotter et al. 2018), or generative-adversarial training (Madras et al. 2018; Zhang, Lemoine, and Mitchell 2018; Celis and Keswani 2019; Xu et al. 2018; Adel et al. 2019).

The closest to our work in this category is the fair robust log-loss predictor of Rezaei et al. (2020), which operates under an iid assumption. That formulation is similar to ours in that it builds a minimax game between a predictor and worst-case approximator of the true distribution. The main difference is that under the iid assumption, the true/false positive rates of target groups can be expressed as a linear constraint on the source data, in which the ground truth label is known. However, in our work, no hard constraint is available for measuring true/false positive rates on target data, because the target true label is unknown under the covariate shift assumption. Thus, we enforce these fairness measures by an expected penalty of the worst-case approximation of the target data.

Covariate Shift

General distribution and domain shift works focus on the joint distribution shift between the training and testing datasets (Daume III and Marcu 2006; Ben-David et al. 2007; Blitzer et al. 2008). Particular assumptions like covariate shift (Shimodaira 2000; Sugiyama, Krauledat, and Müller 2007; Gretton et al. 2009) and label shift (Schölkopf et al. 2012; Lipton, Wang, and Smola 2018; Azizzadenesheli et al.

2019) help quantify the distribution shift using importance weights since they introduce invariance in conditional distributions between the training and testing. Importance weighting methods under covariate shift suffer from high variance and sensitivity to weight estimation methods. It has been shown to often be brittle—providing no finite sample generalization guarantees—even for seemingly benign amounts of shift (Cortes, Mansour, and Mohri 2010). Applying importance weighting to fair prediction has not been broadly investigated and may suffer from a similar issue.

Fairness under perturbation of the attribute has been studied by Awasthi, Kleindessner, and Morgenstern (2020). Lamy et al. (2019) study fair classification when the attribute is subjected to noise according to a mutually contaminated model (Scott, Blanchard, and Handy 2013). Our method works for a general shift on the joint distribution of attribute and features, and does not rely on a particular noise model.

Causal analysis has also been proposed for addressing fairness under dataset shift (Singh et al. 2019). It requires a known causal graph of the data generating process, with a context variable causing the shift, as well as the known separating set of features. Our model makes no assumptions about the underlying structure of the covariates. We assume covariate shift, which relates with causal models when there is no unobserved confounders between covariates and labels. Given a known separating set of features in the causal model under data shift, the covariate shift assumption holds if we use only the separating set of features for prediction. Our model builds on robust classification method of (Liu and Ziebart 2014) under covariate shift, where the target distribution is estimated by an worst-case adversary that maximizes the log-loss while matching the feature statistics under source distribution. Therefore, if we know the separating set of features, we can incorporate them as constraints for the adversary. However, it is usually difficult to know the exact causal model of the data generating process in practice.

Approach

Preliminaries & Notation

We assume a binary classification task $Y, \widehat{Y} \in \{0, 1\}$, where Y denotes the true label, and $\widehat{Y}$ denotes the prediction for a given instance with features $X \in \mathcal{X}$ and group attribute $A \in \{0, 1\}$. We consider $y = 1$ as the privileged class (e.g., an applicant who would repay a loan). Further we assume a given source distribution $(X, A, Y) \sim P_{\text{src}}$ over features, attribute and label and a target distribution $(X, A) \sim P_{\text{trg}}$ over features and attribute only, throughout our paper.

Fairness Definitions Our model seeks to satisfy the group fairness notions of equalized opportunity and odds (Hardt, Price, and Srebro 2016).

Our focus in this paper is equalized opportunity, which requires equal true positive rates across groups, i.e., for a general probabilistic classifier P:

$$P(\widehat{Y}=1|A=1, Y=1) = P(\widehat{Y}=1|A=0, Y=1). \quad (1)$$

Our model can be generalized for equal odds, which in addition to providing an equal true positive rates across groups, also requires equal false positive rates across groups:

$$P(\widehat{Y}=1|A=1, Y=0) = P(\widehat{Y}=1|A=0, Y=0). \quad (2)$$

For Demographic parity (Calders, Kamiran, and Pechenizkiy 2009) which requires equal positive rates across protected groups, i.e $P(\widehat{Y} = 1|A = 1) = P(\widehat{Y} = 1|A = 0)$, our model reduces to a special case, as we later explain.

Covariate Shift In the context of fair prediction, the covariate shift assumption is that the distribution of covariates and group membership can shift between source and target distributions:

$$P_{\text{src}}(x, a, y) = P_{\text{src}}(x, a)P(y|x, a) \quad (3)$$

$$P_{\text{trg}}(x, a, y) = P_{\text{trg}}(x, a)P(y|x, a). \quad (4)$$

Note that we do not assume how the sensitive group membership a is correlated with other features x. If causal structure between the features and labels were known, as assumed in Singh et al. (2019), we could also incorporate a hidden or latent covariate shift assumption. For example, given that there is no unobserved confounder between the covariate and the labels, if $h = \Phi(x, a)$ represent the separating set of features, we can assume $P_{\text{src}}(y|h = \Phi(x, a)) = P_{\text{trg}}(y|h = \Phi(x, a))$ and use $\Phi(x, a)$ instead of (x, a) in our formulation. In this paper, we still use (x, a) to represent the covariates for simplicity.

Importance Weighting A standard approach for addressing covariate shift is to reweight the source data to represent the target distribution (Sugiyama, Krauledat, and Müller 2007). A desired statistic $f(x, a, y)$ of the target distribution can be obtained using samples from the source distribution $(x_i, a_i, y_i)_{i=1:n}$:

$$\mathbb{E}_{\substack{x,a\sim P_{\text{trg}},\\ y|x,a\sim P}}[f(X, A, Y)] \approx \sum_{i=1}^{n} \frac{P_{\text{trg}}(x_i, a_i)}{P_{\text{src}}(x_i, a_i)} f(x_i, a_i, y_i).$$

As long as the source distribution has support for the entire target distribution (i.e., $P_{\text{trg}}(x, a) > 0 \implies P_{\text{src}}(x, a) > 0$), this approximation is exact asymptotically as $n \to \infty$. However, the approximation is only guaranteed to have bounded error for finite n if the source distribution's support for target distribution samples is lower bounded (Cortes, Mansour, and Mohri 2010): $\mathbb{E}_{P_{\text{trg}}(x,a)}[P_{\text{trg}}(X, A)/P_{\text{src}}(X, A)] < \infty$. Unfortunately, this requirement is fairly strict and will not be satisfied even under common and seemingly benign amounts of shift. For example, if source and target samples are drawn from Gaussian distributions with equal (co-)variance, but slightly different means, it is not satisfied.

Robust Log Loss Classification under Covaraite Shift We base our method on the robust approach of Liu and Ziebart (2014) for covariate shift, which addresses this fragility of reweighting methods. The predictor $\mathbb{P}$ minimizes the log loss on a worst-case approximation of the target distribution provided by an adversary $\mathbb{Q}$ that maximizes the log loss while matching the feature statistics of the source distribution:

$$\begin{aligned}&\min_{\mathbb{P}(y|\mathbf{x})\in\Delta}\max_{\mathbb{Q}(y|\mathbf{x})\in\Delta\cap\Xi}\mathbb{E}_{P_{\text{trg}}(\mathbf{x})\mathbb{Q}(y|\mathbf{x})}[-\log\mathbb{P}(Y|\mathbf{X})]\\&=\max_{\mathbb{P}(y|\mathbf{x})\in\Delta\cap\Xi}H_{P_{\text{trg}}(\mathbf{x})\mathbb{P}(y|\mathbf{x})}(Y|\mathbf{X}),\end{aligned}\tag{5}$$

where a moment-matching constraint set $\Xi = \{\mathbb{Q}\,|\,\mathbb{E}_{P_{\text{src}}(\mathbf{x})\mathbb{Q}(y|\mathbf{x})}[\phi(\mathbf{X},Y)] = \mathbb{E}_{P_{\text{src}}(\mathbf{x},y)}[\phi(\mathbf{X},Y)]\}$ on source data is enforced with $\phi(\mathbf{x},y)$ denoting the feature function, and Δ denoting the conditional probability simplex. The saddle point solution under these assumptions is $\mathbb{P}=\mathbb{Q}$ which reduces the formulation to maximizing the target distribution conditional entropy (H) while matching feature statistics of the source distribution. The predictor under these assumptions has the following parametric form:

$$\mathbb{P}_\theta(y|\mathbf{x}) = e^{\frac{P_{\text{src}}(\mathbf{x})}{P_{\text{trg}}(\mathbf{x})}\theta^\top\phi(\mathbf{x},y)}\Big/\sum_{y'\in\mathcal{Y}}e^{\frac{P_{\text{src}}(\mathbf{x})}{P_{\text{trg}}(\mathbf{x})}\theta^\top\phi(\mathbf{x},y')},\tag{6}$$

where the Lagrange multipliers θ are obtained by maximizing the target distribution log likelihood in the dual optimization problem.

Robust Log Loss for Fair Classification (IID) The same robust log loss approach has been employed by Rezaei et al. (2020) for fair classification under the iid assumption ($P_{\text{trg}} = P_{\text{src}}$), where both objective and feature constraints are evaluated on the same training set. Since the true label is available during training, fairness can be enforced as a set of linear constraints on predictor $\mathbb{P}$, which yields a parametric dual form.

In contrast, under the non-iid assumption, the desired fairness on target cannot be directly inferred by enforcing constraints on the source. Additionally, for true/false positive rates defining equalized opportunity (and odds), no linear constraints defined according to the target data are due to the absence of ground truth target labels. Thus, we seek fairness on target data by augmenting the objective in (5) with a penalty incurred by the worst-case approximator of the target $\mathbb{Q}$. In our formulation, the saddle point solution is no longer simple (i.e., $\mathbb{P}\neq\mathbb{Q}$), and no parametric form solution is available.

Formulation

Our formulation seeks a robust and fair predictor under the covariate shift assumption by playing a minimax game augmented by a fairness penalty between a minimizing predictor against a worst-case approximator of the target distribution that matches the feature statistics of the source. We assume the availability of a set of labeled examples from source $\{\mathbf{x}_i,a_i,y_i\}_{i=1}^n \sim P_{\text{src}}(\mathbf{x},a,y)$ and unlabeled examples from target distribution, $\{\mathbf{x}_i,a_i\}_{i=1}^m \sim P_{\text{trg}}(\mathbf{x},a)$ during training.

Definition 1. *The* **Fair Robust Log-Loss Predictor under Covariate Shift**, *$\mathbb{P}$ minimizes the worst-case log loss with an added fairness penalty, while an approximator $\mathbb{Q}$ constrained to reflect source distribution statistics (denoted by set Ξ) maximizes the same fairness-penalized loss:*

$$\begin{aligned}\min_{\mathbb{P}\in\Delta}\max_{\mathbb{Q}\in\Delta\cap\Xi\cap\Gamma}&\mathbb{E}_{P_{trg}(\mathbf{x},a)\mathbb{Q}(y|\mathbf{x},a)}[-\log\mathbb{P}(Y|\mathbf{x},a)]\\&+\mu\,\mathbb{E}_{P_{trg}(\mathbf{x},a)\mathbb{Q}(y'|\mathbf{x},a)\mathbb{P}(y|\mathbf{x},a)}[f(A,Y',Y)]\end{aligned}\tag{7}$$

such that:

$$\begin{aligned}&\Xi(\mathbb{Q}):\mathbb{E}_{\substack{P_{src}(\mathbf{x},a)\\\mathbb{Q}(y|\mathbf{x},a)}}[\phi(\mathbf{X},Y)] = \mathbb{E}_{P_{src}(\mathbf{x},a,y)}[\phi(\mathbf{X},Y)]\text{ and}\\&\forall k\in\{0,1\},\\&\Gamma(\mathbb{Q}):\mathbb{E}_{\substack{P_{trg}(\mathbf{x},a)\\\mathbb{Q}(y|\mathbf{x},a)}}[g_k(A,Y)] = \underbrace{\mathbb{E}_{\substack{P_{trg}(\mathbf{x},a)\\\widehat{P}_{trg}(y|\mathbf{x},a)}}[g_k(A,Y)]}_{\widetilde{g}_k},\end{aligned}$$

where ϕ is the feature function, μ is the fairness penalty weight, $g_k(.,.)$ is a selector function for group k according to the fairness definition, i.e., for equalized opportunity: $g_k(A,Y) = \mathbb{I}(A=k\wedge Y=1)$, $\widetilde{g}_k$ the estimated group density on target, and $f(.,.,.)$ is a weighting function of the mean score difference between the two groups:

$$f(A,Y,\widehat{Y}) = \begin{cases}\frac{1}{\widetilde{g}_1} & \textit{if } g_1(A,Y)\wedge\mathbb{I}(\widehat{Y}=1)\\ -\frac{1}{\widetilde{g}_0} & \textit{if } g_0(A,Y)\wedge\mathbb{I}(\widehat{Y}=1)\\ 0 & \textit{otherwise.}\end{cases}\tag{8}$$

The Γ constraint enforces $\mathbb{Q}$ to be consistent with the marginal probability of the groups on target ($\widetilde{g}_k$) for equalized opportunity (and odds). This marginal probability is unknown, since the true label Y on target is unavailable. Thus, we estimate these marginal probabilities by employing the robust model (6) as $\widehat{P}_{\text{trg}}(y|\mathbf{x},a)$ in (7) to first guess the labels under covariate shift ignoring fairness ($\mu=0$). We penalize the expected difference in true (or false) positive rate of groups in target according to our worst-case approximation of each example being positive label. This needs to be measured on the entire target example set and requires batch gradient updates to enforce.

Our formulation is flexible for all three mentioned definitions of group fairness. For equalized odds, a second penalty term for false positive rates ($g_k(A,Y) = \mathbb{I}(A=k\wedge Y=0)$) is required and the corresponding marginal matching constraint in Γ needs to be added. For demographic parity ($g_k(A,Y) = \mathbb{I}(A=k)$), because the group definition is independent of the true label, the target groups are fully known and the fairness penalty reduces to a linear constraint of $\mathbb{P}$ on target. In this special case, there is no need for the Γ constraint and μ can be treated as a dual variable for the linear fairness constraint. This reduces to the truncated logistic classifier of (Rezaei et al. 2020) with the exception that the fairness constraint is formed on the target data.

We obtain the following solution for the predictor $\mathbb{P}$ by leveraging strong minimax duality (Topsøe 1979; Grünwald and Dawid 2004) and strong Lagrangian duality (Boyd and Vandenberghe 2004).

Theorem 1. *The* **Fair Robust Log-Loss Predictor under Covariate Shift** (7) *for equalized opportunity with a given fairness penalty parameter, μ, can be obtained by solving:*

$$\begin{aligned}&\log\frac{1-\mathbb{P}(y|\mathbf{x},a)}{\mathbb{P}(y|\mathbf{x},a)}+\mu\,\mathbb{E}_{\mathbb{P}(y'|\mathbf{x},a)}\left[f(a,y,Y')\right]\\&+\frac{P_{src}(\mathbf{x},a)}{P_{trg}(\mathbf{x},a)}\theta^{\mathrm{T}}\left(\phi(\mathbf{x},y=1)-\phi(\mathbf{x},y=0)\right)\\&+\sum_{k\in 0,1}\lambda_k g_k(a,y)=0,\end{aligned}\tag{9}$$

where θ and λ are the dual Lagrange multipliers for moment matching constraints (Ξ) and target marginal matching (Γ) respectively.

Given the solution $\mathbb{P}^*$ obtained above, for $\mathbb{Q}$ to be in equilibrium (given θ) it suffices to choose $\mathbb{Q}$ such that:

$$\mathbb{Q}(y|\mathbf{x}, a) = \frac{\mathbb{P}^*(y|\mathbf{x}, a)}{1 - \mu f(a, y, y) + \mu f(a, y, y)\mathbb{P}^{*2}(y|\mathbf{x}, a)} \tag{10}$$

where $0 \leq \mathbb{Q}(y|x, a) \leq 1$.

Due to monotonicity, $\mathbb{P}$ in (9) is efficiently found using a binary-search in the simplex. For proofs and further details, we refer the interested reader to the appendix.

Enforcing Fairness

Our model approximates the fairness violation on the target distribution by the incurred fairness penalty measured on the robust target approximator ($\mathbb{Q}$). We seek optimal μ by finding the zero-point of approximated fairness penalty, i.e., $\mathbb{E}_{P_{\text{trg}}(\mathbf{x},a)\mathbb{Q}(y'|\mathbf{x},a)\mathbb{P}(y|\mathbf{x},a)}[f(A, Y', Y)]$ in (9). Under a mild assumption that $\mathbb{Q}$ is sufficiently constrained by source feature statistics and marginal probability of the target groups, the approximated fairness violation by $\mathbb{Q}$ is monotone in the proximity of the zero point. Thus we can find the exact zero point by a binary-search in a neighborhood around zero point.

Learning

For a given fairness penalty μ, our model seeks to learn the dual parameters θ and λ, such that the worst-case log-loss approximation ($\mathbb{Q}_{\theta,\mu}$) matches the sample feature statistic from the source distribution and the marginal group probability on the target set. Given θ^*, λ^* the solution of (9) obtains the optimal fair predictor $\mathbb{P}^*_{\theta^*,\mu^*}$ which is robust against the sampling shift.

We employ L2 regularization on θ parameters to improve the generalizability of our model. This corresponds to relaxing our feature matching constraints by a convex norm. We employ a batch gradient algorithm to learn our model parameters. We perform a joint gradient optimization that updates the gradient of θ (which depends on the true label) from the source data and λ (which does not depend on true label) from the target batch at each step. Note that we find the solution to the dual objective of (9) by gradient-only optimization without requiring the explicit calculation of the objective on the target dataset. The gradient optimization converges to the global optimum because the dual objective is convex in θ and λ.

Given an optimal $\mathbb{Q}^*$ from (10), a set of labeled source training samples $\widetilde{P}_{\text{src}}(\mathbf{x}, a, y)$ and unlabeled target samples $\widetilde{P}_{\text{trg}}(\mathbf{x}, a)$, the gradient of our model parameters is calculated as follows:

$$\nabla_\theta \mathcal{L}^{\text{trg}}(\mathbb{P}, \mathbb{Q}, \theta, \lambda) = \mathbb{E}_{\substack{\widetilde{P}_{\text{src}}(\mathbf{x},y) \\ \mathbb{Q}(\widehat{y}|\mathbf{x},a)}} \phi(\mathbf{X}, \widehat{Y}) - \mathbb{E}_{\widetilde{P}_{\text{src}}(\mathbf{x},a,y)} \phi(\mathbf{X}, Y)$$
$$\nabla_{\lambda_{a'}} \mathcal{L}^{\text{trg}}(\mathbb{P}, \mathbb{Q}, \theta, \lambda) = \mathbb{E}_{\substack{\widetilde{P}_{\text{trg}}(\mathbf{x},a) \\ \mathbb{Q}(y|\mathbf{x},a)}} [g_k(A, Y)] - \widetilde{g}_k. \tag{11}$$

Note that although the gradient of θ can be updated stochastically, the λ gradient update relies on calculating the $\mathbb{Q}$ marginal for each group on a target batch. This process is described in detail in Algorithm 1.

Algorithm 1: Batch gradient update for Fair Robust Log-Loss learning under Covariate Shift

```
Input: Datasets $\widetilde{P}_{\text{src}} = \{\mathbf{x}_i, a_i, y_i\}_{i=1}^n$,
       $\widetilde{P}_{\text{trg}} = \{\mathbf{x}_i, a_i\}_{i=1}^m$, ratio $\frac{P_{\text{src}}}{P_{\text{trg}}}$, feature function
       $\phi(\mathbf{x}, y)$, decaying learning rate $\eta_t$
Output: $\theta^*, \lambda^*$
$\theta \leftarrow$ random initialization ;
$\lambda_a \leftarrow 0$ ;
repeat
    Compute $\mathbb{P}, \mathbb{Q}$ for all source dataset examples by
      finding solution to (9) and (10)
    $\nabla_\theta \mathcal{L} \leftarrow \frac{1}{n}\sum_{i=1}^n \phi(\mathbf{x}_i, y_i) -$
      $\frac{1}{n}\sum_{i=1}^n \sum_{y\in\mathcal{Y}} \mathbb{Q}(y|\mathbf{x}_i, a)\phi(\mathbf{x}_i, y)$ ;
    $\nabla_{\lambda_k} \mathcal{L} \leftarrow \widetilde{g}_k - \frac{1}{m}\sum_{i=1}^m \sum_{y\in\mathcal{Y}} \mathbb{Q}(y|\mathbf{x}, a)g_k(a, y)$
      ;
    $\theta \leftarrow \eta_t(\nabla_\theta \mathcal{L} + C\nabla\|\theta\|)$
    $\lambda \leftarrow \eta_t(\nabla_\lambda \mathcal{L})$ ;
until convergence;
```

Experiments

We demonstrate the effectiveness of our method on benchmark data sets containing sensitive features. We compare with two sets of baselines: (1) previous methods enforcing fairness constraints under the iid assumption and (2) covariate shift classification models that do not account for fairness. Our experiments shows that our approach outperform the baselines by achieving smaller fairness constraint violations without sacrificing predictive performance.

We evaluate our method against baselines on four datasets:

- The COMPAS criminal recidivism risk assessment dataset (Larson et al. 2016). The task is to predict recidivism of a defendant based on criminal history.
- UCI Drug dataset (Fehrman et al. 2017). The task is to classify type of drug consumer by personality and demongraphics.
- UCI Arrhythmia dataset (Dheeru and Karra Taniskidou 2017). The task is to distinguish between the presence and absence of cardiac arrhythmia.
- UCI German dataset (Dheeru and Karra Taniskidou 2017). The task it to classify good and bad credit according to personal information and credit history.

Table 1 summarizes the characteristics of each of these datasets.

We create biased samplings by modeling a general shift between the distribution of the covariates in source and target, i.e., $P_{\text{src}}(\mathbf{x}, a) \neq P_{\text{trg}}(\mathbf{x}, a)$.We take the following steps to create covariate shift on the normalized dataset:

1. We apply principal component analysis (PCA) to retrieve the first principal component $\mathcal{C}$ from features.
2. We estimate the mean and standard deviation of P as $\mu(\mathcal{C})$ and $\sigma(\mathcal{C})$.
3. We choose parameters m, s and set a normal distribution as $D_s(\mu(\mathcal{C}) + m, \frac{\sigma(\mathcal{C})}{s})$.
4. We sample from the data with probability in proportion to D_s to construct the source data distribution.
5. We sample from the data with probability in proportion to $D_t(\mu(\mathcal{C}), \sigma(\mathcal{C}))$ to construct the target data distribution.

Dataset	n	Features	Attribute
COMPAS	6,167	10	Race
German	1,000	20	Gender
Drug	1,885	11	Race
Arrhythmia	452	279	Gender

Table 1: Dataset characteristics.

Baseline methods

We evaluate the performance of our model in terms of the trade-off between incurring log loss and fairness violation under various degrees of covariate shift. We focus on using equalized opportunity as our fairness definition. We compare against the following baselines [1]:

- **Logistic Regression** (LR) is the standard logistic regression trained on source data: it ignores both covariate shift and fairness assumptions.
- **Robust Bias-Aware Log Loss Classifier** (RBA) (Liu and Ziebart 2014) robustly minimizes the worst-case log loss on the target distribution while matching feature statistics of the source; it accounts for the covariate shift but ignores fairness.
- **Sample Reweighted Logistic Regression** (LR_IW) (Shimodaira 2000) minimizes the reweighted log loss on the source data, according to the importance weighting ratio: it only accounts for the covariate shift.
- **Source Fair Logistic Regression** (FAIRLR) is the method of (Rezaei et al. 2020) that optimizes worst-case log loss subject to fairness as linear constraints on labeled source data. It accounts for fairness but ignores the covariate shift. We choose this baseline as a candidate in-processing method under IID assumptions. This baseline has been shown to have performance on the Pareto frontier of the prediction-fairness trade-off against other in-processing methods such as Zafar et al. (2017c) and cost-sensitive reduction approach of Agarwal et al. (2018). However, the underlying model selection assumption for all of these existing methods is violated under the covariate shift setting.
- **Sample Reweighted Fair Logistic Regression** (FAIRLR_IW) the fairLR method augmented with importance weighting ratios to the log loss in objective. This baseline account for both fairness and covariate shift.

Setup We repeat our sampling procedure for each dataset ten times and report the average log loss and the average difference of equalized opportunity (DEO): $|\mathbb{P}(\widehat{Y} = 1|A = 1, Y = 1) - \mathbb{P}(\widehat{Y} = 1|A = 0, Y = 1)|$ of our predictor on the target dataset.

Unfortunately since the target distribution is assumed to be unavailable for this problem, properly obtaining optimal regularization via cross validation is not possible. We select the L2 regularization parameter by choosing the best C from $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ under the IID assumption. We use first-order features for our implementation, i.e $\phi(\mathbf{x}, y) = [x_1 y, x_2 y, \dots x_m y]^\top$, where m is the size of features.

Assuming sufficient expressive feature constraints, $\mathbb{Q}$ is always sufficiently constrained and thus under mild assumption that the approximated DEO by $\mathbb{Q}$ remains monotone in relatively small intervals of $[\mu, \mu + \epsilon]$ we first perform a line search on $\mu \in [-1, 1]$ with intervals of $\epsilon = .1$ to find the zero crossing regions, and then we find the exact zero point of the approximated violation efficiently by binary search.

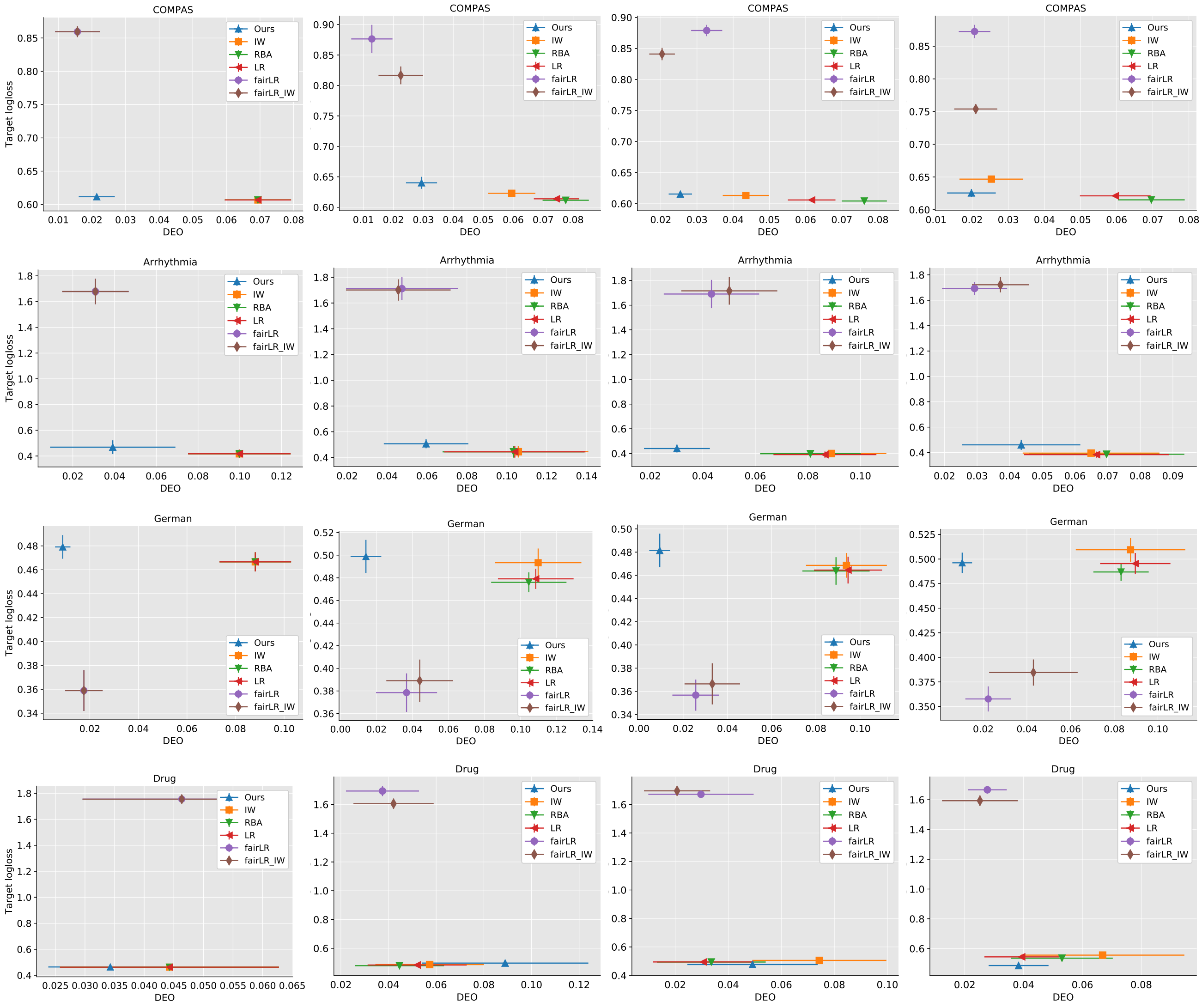
Results

Figure 2 shows our experiment results on iid target data and three shifted samplings.

On the COMPAS dataset, our method's logloss and DEO remains relatively unchanged with different shift settings. Although FAIRLR methods provides better fairness, their logloss is relatively high. As the sampling variance intensifies, the importance weighting methods' DEO also gets better. In all samplings our method remains on the frontier of logloss and DEO.On the Arrhythmia dataset, our method's logloss remains relatively low, only slightly worse than unfair methods. Whereas compared to the fair methods, ours provide competitive DEO in the overlapping range, with little trade-off in logloss. On this dataset our method is less sensitive to shift in variance than the shift in the mean. When the sampling shift is only due to the variance (third column), our method provides lowest logloss and lowest DEO. On the German dataset our method provides the lowest DEO on all shifted samplings, with similar logloss to unfair methods and higher log loss than FAIRLR methods. On the Drug dataset our method's log loss remains low, while on DEO dimension it lies competitive with other baselines. Similar to Arrhythmia, our method fairness performance is better under variance-based shifts on this dataset. As the shift intensifies, our method's log loss remains lower than all other baselines and the DEO nears other fair methods.

In summary, on the COMPAS and Arrhythmia dataset our method provides better trade-off of lower DEO and low logloss. On German we have the best fairness score while having similar logloss to unfair methods. On the Drug dataset, our method nears the best trade-off on the sampling with the strongest shift.

[1] We do not include post-processing baselines since the available codes (Hardt, Price, and Srebro 2016) only work for equalized odds while our experiments focus on equalized opportunity.



$m = 0, s = 1$ (IID) $m = 0.5, s = 1$ $m = 0, s = 1.5$ $m = 0.5, s = 1.5$

Figure 2: Average *log loss* versus average *difference of equalized opportunity* (DEO) on target samples. The bar is the 95% confidence interval on ten random biased samplings of covariates ($P_{\text{src}}(x, a) \neq P_{\text{trg}}(x, a)$) based on the first principal component.

Conclusions

In this paper, we developed a novel adversarial approach for seeking fair decision making under covariate shift. In contrast with importance weighting methods, our approach is designed to operate appropriately even when portions of the shift between source and target distributions are extreme. The key technical challenge we address is the lack of labeled target datapoints, making target fairness assessment challenging. We instead propose to measure fairness against an adversary that is constrained by source data properties. We incorporate fairness as a weighted penalty and tune the weighted penalty to provide fairness against the adversary. More extensive evaluation on naturally-biased datasets and generalization of this approach to decision problems beyond binary classification are both important directions for future work.

Broader Impact

Fairness considerations are increasingly important for machine learning systems applied to key social applications. However, the standard assumptions of statistical machine learning, such as iid training and testing data, are often violated in practice. This work offers an approach for robustly seeking fair decisions in such settings and could be of general benefit to individuals impacted by alternative systems that are either oblivious or brittle to these broken assumptions. However, this work also makes a covariate shift assumption instead of accounting for more specific causal relations that may generate the shift. Practitioners should be aware of the specific assumptions made by this paper.

References

Adel, T.; Valera, I.; Ghahramani, Z.; and Weller, A. 2019. One-network Adversarial Fairness. In *AAAI*.

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. M. 2018. A Reductions Approach to Fair Classification. In *ICML*.

Awasthi, P.; Kleindessner, M.; and Morgenstern, J. 2020. Equalized odds postprocessing under imperfect group information. In *International Conference on Artificial Intelligence and Statistics*, 1770–1780. PMLR.

Azizzadenesheli, K.; Liu, A.; Yang, F.; and Anandkumar, A. 2019. Regularized learning for domain adaptation under label shifts. *arXiv preprint arXiv:1903.09734* .

Barocas, S.; Hardt, M.; and Narayanan, A. 2017. Fairness in machine learning. *NIPS Tutorial* .

Bechavod, Y.; and Ligett, K. 2017. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044* .

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

Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Wortman, J. 2008. Learning bounds for domain adaptation. In *Advances in neural information processing systems*, 129–136.

Boyd, S.; and Vandenberghe, L. 2004. *Convex optimization*. Cambridge university press.

Calders, T.; Kamiran, F.; and Pechenizkiy, M. 2009. Building classifiers with independency constraints. In *ICDMW '09*.

Calmon, F.; Wei, D.; Vinzamuri, B.; Natesan Ramamurthy, K.; and Varshney, K. R. 2017. Optimized Pre-Processing for Discrimination Prevention. In *NeurIPS*.

Carter, C.; and Catlett, J. 1987. Assessing credit card applications using machine learning. *IEEE Expert* .

Celis, L. E.; Huang, L.; Keswani, V.; and Vishnoi, N. K. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *ACM FAT**.

Celis, L. E.; and Keswani, V. 2019. Improved Adversarial Learning for Fair Classification. *arXiv preprint* .

Chang, L. 2006. Applying data mining to predict college admissions yield: A case study. *NDIR* .

Chierichetti, F.; Kumar, R.; Lattanzi, S.; and Vassilvitskii, S. 2017. Fair clustering through fairlets. In *Advances in Neural Information Processing Systems*, 5029–5037.

Cortes, C.; Mansour, Y.; and Mohri, M. 2010. Learning bounds for importance weighting. In *Advances in neural information processing systems*, 442–450.

Cotter, A.; Jiang, H.; Wang, S.; Narayan, T.; Gupta, M.; You, S.; and Sridharan, K. 2018. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *arXiv preprint* .

Daume III, H.; and Marcu, D. 2006. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research* 26: 101–126.

Del Barrio, E.; Gamboa, F.; Gordaliza, P.; and Loubes, J.-M. 2018. Obtaining fairness using optimal transport theory. *arXiv preprint* .

Dheeru, D.; and Karra Taniskidou, E. 2017. UCI Machine Learning Repository. URL http://archive.ics.uci.edu/ml.

Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J. S.; and Pontil, M. 2018. Empirical risk minimization under fairness constraints. In *NeurIPS*.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *ITCS*.

Dwork, C.; Immorlica, N.; Kalai, A. T.; and Leiserson, M. 2017. Decoupled classifiers for fair and efficient machine learning. *arXiv preprint arXiv:1707.06613* .

Fehrman, E.; Muhammad, A. K.; Mirkes, E. M.; Egan, V.; and Gorban, A. N. 2017. The five factor model of personality and evaluation of drug consumption risk. In *Data science*, 231–242. Springer.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *ACM SIGKDD*.

Goel, N.; Yaghini, M.; and Faltings, B. 2018. Non-discriminatory machine learning through convex fairness criteria. In *AAAI*.

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.

Grünwald, P. D.; and Dawid, A. P. 2004. Game Theory, Maximum Entropy, Minimum Discrepancy, and Robust Bayesian Decision Theory. *Annals of Statistics* 32.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51(5): 1–42.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *NeurIPS*.

Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; and Roth, A. 2016. Fair learning in markovian environments. *arXiv preprint arXiv:1611.03071* .

Kabakchieva, D. 2013. Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies* 13(1).

Kamiran, F.; and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33(1).

Kamishima, T.; Akaho, S.; and Sakuma, J. 2011. Fairness-aware learning through regularization approach. In *ICDMW*.

Lamy, A.; Zhong, Z.; Menon, A. K.; and Verma, N. 2019. Noise-tolerant fair classification. In *Advances in Neural Information Processing Systems*, 294–306.

Larson, J.; Mattu, S.; Kirchner, L.; and Angwin, J. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica* .

Lipton, Z. C.; Wang, Y.-X.; and Smola, A. 2018. Detecting and correcting for label shift with black box predictors. *arXiv preprint arXiv:1802.03916* .

Liu, A.; and Ziebart, B. 2014. Robust Classification Under Sample Selection Bias. In *NeurIPS*.

Lohr, S. 2013. Big data, trying to build better workers. *The New York Times* 21.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning adversarially fair and transferable representations. *arXiv preprint* .

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* .

Menon, A. K.; and Williamson, R. C. 2018. The cost of fairness in binary classification. In *ACM FAT**.

Moses, L. B.; and Chan, J. 2014. Using big data for legal and law enforcement decisions: Testing the new tools. *UNSWLJ* .

Obermeyer, Z.; and Emanuel, E. J. 2016. Predicting the future—big data, machine learning, and clinical medicine. *The New England Journal of Medicine* 375(13).

O'Neil, C. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Pedreshi, D.; Ruggieri, S.; and Turini, F. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 560–568.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On Fairness and Calibration. In *NeurIPS*.

Quadrianto, N.; and Sharmanska, V. 2017. Recycling privileged learning and distribution matching for fairness. In *NeurIPS*.

Rezaei, A.; Fathony, R.; Memarrast, O.; and Ziebart, B. 2020. Fairness for Robust Log Loss Classification. In *AAAI*.

Schölkopf, B.; Janzing, D.; Peters, J.; Sgouritsa, E.; Zhang, K.; and Mooij, J. 2012. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471* .

Scott, C.; Blanchard, G.; and Handy, G. 2013. Classification with asymmetric label noise: Consistency and maximal denoising. In *Conference On Learning Theory*, 489–511.

Shaw, M. J.; and Gentry, J. A. 1988. Using an expert system with inductive learning to evaluate business loans. *Financial Management* .

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.

Shipp, M. A.; Ross, K. N.; Tamayo, P.; Weng, A. P.; Kutok, J. L.; Aguiar, R. C.; Gaasenbeek, M.; Angelo, M.; Reich, M.; Pinkus, G. S.; et al. 2002. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine* 8(1).

Singh, H.; Singh, R.; Mhasawade, V.; and Chunara, R. 2019. Fair Predictors under Distribution Shift. *arXiv preprint arXiv:1911.00677* .

Sugiyama, M.; Krauledat, M.; and Müller, K.-R. 2007. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research* 8(May): 985–1005.

Topsøe, F. 1979. Information-theoretical optimization techniques. *Kybernetika* 15(1): 8–27.

Verma, S.; and Rubin, J. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, 1–7. IEEE.

Woodworth, B.; Gunasekar, S.; Ohannessian, M. I.; and Srebro, N. 2017. Learning Non-Discriminatory Predictors. In *COLT*.

Xu, D.; Yuan, S.; Zhang, L.; and Wu, X. 2018. FairGAN: Fairness-aware generative adversarial networks. In *IEEE Big Data*.

Yang, K.; Qinami, K.; Fei-Fei, L.; Deng, J.; and Russakovsky, O. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 547–558.

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

Zafar, M. B.; Valera, I.; Rodriguez, M.; Gummadi, K.; and Weller, A. 2017b. From parity to preference-based notions of fairness in classification. In *NeurIPS*.

Zafar, M. B.; Valera, I.; Rogriguez, M. G.; and Gummadi, K. P. 2017c. Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning Fair Representations. In *ICML*.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *AIES*.

Supplementary Materials

Proof of Theorems

Proof of Theorem 1. First, we incorporate the source feature-matching and target marginal matching constraint and the normalization constraint for $\mathbb{P}$ with Lagrangian multipliers θ, λ and $Z_{\mathbb{P}}$ into the objective. To generalize for equalized odds we extend our fairness-related variables to $\mathbb{R}^{2\times 1}$ vectors to include false positive rates as well, i.e $\boldsymbol{\mu} = [\mu^{\text{tp}}; \mu^{\text{fp}}]$, $\boldsymbol{\lambda}_k = [\lambda_k^{\text{tp}}; \lambda_k^{\text{fp}}]$, $\mathbf{g_k}(A, Y) = [g_k^{\text{tp}}; g_k^{\text{fp}}] = [\mathbb{I}(A = k, Y = 1); \mathbb{I}(A = k, Y = 0)]$, and $\mathbf{f} = [f^{\text{tp}}; f^{\text{fp}}]$:

$$f^{\text{tp}}(A, Y, \widehat{Y}) = \begin{cases} \frac{1}{\widetilde{g}_1^{\text{tp}}} & \text{if } g_1^{\text{tp}}(A, Y) \wedge \mathbb{I}(\widehat{Y}=1) \\ -\frac{1}{\widetilde{g}_0^{\text{tp}}} & \text{if } g_0^{\text{fp}}(A, Y) \wedge \mathbb{I}(\widehat{Y}=1) \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

$$f^{\text{fp}}(A, Y, \widehat{Y}) = \begin{cases} \frac{1}{\widetilde{g}_1^{\text{fp}}} & \text{if } g_1^{\text{fp}}(A, Y) \wedge \mathbb{I}(\widehat{Y}=1) \\ -\frac{1}{\widetilde{g}_0^{\text{fp}}} & \text{if } g_0^{\text{fp}}(A, Y) \wedge \mathbb{I}(\widehat{Y}=1) \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

The dual objective is:

$$\begin{aligned} &\min_{\mathbb{P}} \max_{\mathbb{Q}\in\Delta, Z_{\mathbb{P}}} \min_{\theta,\lambda} L(\mathbb{P}, \mathbb{Q}, \theta, \lambda, Z_{\mathbb{P}}) \overset{(a)}{=} \min_{\theta,\lambda} \min_{\mathbb{P}} \max_{\mathbb{Q}\in\Delta, Z_{\mathbb{P}}} \sum_{\mathbf{x},a} P_{\text{trg}}(\mathbf{x}, a) \Bigg(\mathbb{E}_{\mathbb{Q}(y|\mathbf{x},a)} \Bigg[-\log \mathbb{P}(Y|\mathbf{x}, a) \\ &\quad + \boldsymbol{\mu}^\top \mathbb{E}_{\mathbb{P}(y'|\mathbf{x},a)} \left[\mathbf{f}(a, Y, Y')|\mathbf{x}, a\right] + \frac{P_{\text{src}}(\mathbf{x}, a)}{P_{\text{trg}}(\mathbf{x}, a)} \theta^{\mathrm{T}} \big(\phi(\mathbf{x}, Y) - \underbrace{\mathbb{E}_{P_{\text{src}}(y''|\mathbf{x},a)} \left[\phi(\mathbf{x}, Y'')|\mathbf{x}, a\right]}_{\widetilde{\boldsymbol{\phi}}} \big) \\ &\quad + \sum_{k\in\{0,1\}} \boldsymbol{\lambda}_k^\top (\mathbf{g}_k(a, y) - \widetilde{\mathbf{g}}_k)|\mathbf{x}, a \Bigg] + Z_{\mathbb{P}}(\mathbf{x}, a) \Bigg(\sum_y \mathbb{P}(y|\mathbf{x}, a) - 1 \Bigg) \Bigg) \end{aligned} \tag{14}$$

$$\begin{aligned} &\overset{(b)}{=} \min_{\theta,\lambda} \max_{Z_{\mathbb{P}}} -\theta^\top \widetilde{\boldsymbol{\phi}} - \sum_{k\in\{0,1\}} \boldsymbol{\lambda}_k^\top \widetilde{\mathbf{g}}_k + \sum_{\mathbf{x},a} P_{\text{trg}}(\mathbf{x}, a) \Bigg(\min_{\mathbb{P}(.|\mathbf{x},a)} \max_y - \log \mathbb{P}(y|\mathbf{x}, a) + \boldsymbol{\mu}^\top \mathbb{E}_{\mathbb{P}(y'|\mathbf{x},a)} \left[\mathbf{f}(a, y, Y')|\mathbf{x}, a\right] \\ &\quad + \frac{P_{\text{src}}(\mathbf{x}, a)}{P_{\text{trg}}(\mathbf{x}, a)} \theta^{\mathrm{T}} \phi(\mathbf{x}, y) + \sum_{k\in\{0,1\}} \boldsymbol{\lambda}_k^\top \mathbf{g}_k(a, y) + Z_{\mathbb{P}}(\mathbf{x}, a) \left(\sum_y \mathbb{P}(y|\mathbf{x}, a) - 1 \right) \Bigg) \end{aligned} \tag{15}$$

$$\overset{(c)}{\Longrightarrow} - \log \mathbb{P}(y|\mathbf{x}, a) + \boldsymbol{\mu}^\top \mathbb{E}_{\mathbb{P}(y'|\mathbf{x},a)} \left[\mathbf{f}(a, y, Y')\right] + \frac{P_{\text{src}}(\mathbf{x}, a)}{P_{\text{trg}}(\mathbf{x}, a)} \theta^{\mathrm{T}} \phi(\mathbf{x}, y) + \sum_{k\in\{0,1\}} \boldsymbol{\lambda}_k^\top \mathbf{g}_k(a, y) = C(\mathbf{x}, a) \quad \forall y \in \mathcal{Y} \tag{16}$$

for a normalization term $C(\mathbf{x}, a)$; C must be chosen so that: $\sum_y \mathbb{P}(y|\mathbf{x}, a) = 1$. Strong duality is used in step (a) since $L(\mathbb{P}, \mathbb{Q}, \theta)$ is convex in $\mathbb{P}$, convex in θ, and concave (affine) in $\mathbb{Q}$. Since the function is affine in terms of $\mathbb{Q}$, a solution must exist at the extreme point, i.e., $y = 0$ or $y = 1$, the optimization reduces to choosing from those extreme values in step (b). For $\mathbb{P}$ to be minimal, the values for each choice of y must be equal, as shown in step (c).

For binary classification $y \in \{0, 1\}$, the normalization term can be eliminated,and (16) simplifies to:

$$\log \frac{1 - \mathbb{P}(y = 1|\mathbf{x}, a)}{\mathbb{P}(y = 1|\mathbf{x}, a)} + \boldsymbol{\mu}^\top \mathbb{E}_{\mathbb{P}(y'|\mathbf{x},a)} \left[\mathbf{f}(a, y, Y')\right] + \frac{P_{\text{src}}(\mathbf{x}, a)}{P_{\text{trg}}(\mathbf{x}, a)} \theta^{\mathrm{T}} \left(\phi(\mathbf{x}, y = 1) - \phi(\mathbf{x}, y = 0)\right) + \sum_{k\in 0,1} \boldsymbol{\lambda}_k^\top \mathbf{g}_k(a, y) = 0$$

For $\mathbb{Q}$ to be in equilibrium (given θ), the $\mathbb{P}$ obtained above, which we denote as $\mathbb{P}^*$, must be optimal: $L(\mathbb{P}^*, \mathbb{Q}, \theta) \leq L(\mathbb{P}', \mathbb{Q}, \theta), \forall \mathbb{P}'$. Since $L(\mathbb{P}, \mathbb{Q}, \theta)$ is convex in $\mathbb{P}$, it suffices to choose $\mathbb{Q}$ so that:

$$\left.\frac{\partial L(\mathbb{P}, \mathbb{Q}, \theta)}{\partial \mathbb{P}}\right|_{\mathbb{P}=\mathbb{P}^*} = 0 \tag{17}$$

$$\Longrightarrow \; P_{\text{trg}}(\mathbf{x}, a) \left(-\frac{\mathbb{Q}(y|\mathbf{x}, a)}{\mathbb{P}^*(y|\mathbf{x}, a)} + \boldsymbol{\mu}^\top \mathbb{E}_{\mathbb{Q}(y'|\mathbf{x},a)} \left[\mathbf{f}(a, Y', y)|\mathbf{x}, a\right] + Z_{\mathbb{P}}(\mathbf{x}, a) \right) = 0 \quad \forall y \in \mathcal{Y} \tag{18}$$

$$\Longrightarrow \; -\frac{\mathbb{Q}(y|\mathbf{x}, a)}{\mathbb{P}^*(y|\mathbf{x}, a)} + \boldsymbol{\mu}^\top \mathbb{E}_{\mathbb{Q}(y'|\mathbf{x},a)} \left[\mathbf{f}(a, Y', y)|\mathbf{x}, a\right] + Z_{\mathbb{P}}(\mathbf{x}, a) = 0 \quad \forall y \in \mathcal{Y}. \tag{19}$$

where $Z_{\mathbb{P}}(\mathbf{x}, a)$ must be chosen such that $\sum_y \mathbb{Q}^*(y|\mathbf{x}, a) = 1$. For binary classification $y \in \{0, 1\}$ we can expand (19):

$$\implies -\frac{\mathbb{Q}(y|\mathbf{x}, a)}{\mathbb{P}^*(y|\mathbf{x}, a)} + \mu^{\text{tp}}\mathbb{Q}(y|\mathbf{x}, a)f^{\text{tp}}(a, y, y) + \mu^{\text{fp}}(1 - \mathbb{Q}(y|\mathbf{x}, a))f^{\text{fp}}(a, 1 - y, y) + Z_{\mathbb{P}}(\mathbf{x}, a) = 0 \quad \forall y \in \{0, 1\}. \tag{20}$$

$$\implies \mathbb{Q}^*(y|\mathbf{x}, a) = \frac{Z_{\mathbb{P}}(\mathbf{x}, a) + \mu^{\text{fp}} f^{\text{fp}}(a, 1 - y, y)}{\frac{1}{\mathbb{P}^*(y|\mathbf{x},a)} - \mu^{\text{tp}} f^{\text{tp}}(a, y, y) + \mu^{\text{fp}} f^{\text{fp}}(a, 1 - y, y)} \tag{21}$$

$$\sum_y \mathbb{Q}^*(y|\mathbf{x}, a) = 1 \implies \sum_y \frac{Z_{\mathbb{P}}(\mathbf{x}, a) + \mu^{\text{fp}} f^{\text{fp}}(a, 1 - y, y)}{\frac{1}{\mathbb{P}^*(y|\mathbf{x},a)} - \mu^{\text{tp}} f^{\text{tp}}(a, y, y) + \mu^{\text{fp}} f^{\text{fp}}(a, 1 - y, y)} = 1 \tag{22}$$

$$\implies Z_{\mathbb{P}}(\mathbf{x}, a) \sum_y \frac{1}{\frac{1}{\mathbb{P}^*(y|\mathbf{x},a)} - \mu^{\text{tp}} f^{\text{tp}}(a, y, y) + \mu^{\text{fp}} f^{\text{fp}}(a, 1 - y, y)} + \sum_y \frac{\mu^{\text{fp}} f^{\text{fp}}(a, 1 - y, y)}{\frac{1}{\mathbb{P}^*(y|\mathbf{x},a)} - \mu^{\text{tp}} f^{\text{tp}}(a, y, y) + \mu^{\text{fp}} F^{\text{fp}}(a, 1 - y, y)} = 1 \tag{23}$$

$$\implies Z_{\mathbb{P}(\mathbf{x},a)} = \frac{1 - \sum_y \frac{\mu^{\text{fp}} f^{\text{fp}}(a,1-y,y)}{\frac{1}{\mathbb{P}^*(\mathbf{x},a)} - \mu^{\text{tp}} f^{\text{tp}}(a,y,y) + \mu^{\text{fp}} f^{\text{fp}}(a,1-y,y)}}{\sum_y \frac{1}{\frac{1}{\mathbb{P}^*(\mathbf{x},a)} - \mu^{\text{tp}} f^{\text{tp}}(a,y,y) + \mu^{\text{fp}} f^{\text{fp}}(a,1-y,y)}} \tag{24}$$

$$\implies Z_{\mathbb{P}(\mathbf{x},a)} = \frac{1 - \frac{\mu^{\text{fp}} f^{\text{fp}}(a,y=0,y=1)}{\frac{1}{\mathbb{P}^*(y=1|\mathbf{x},a)} - \mu^{\text{tp}} f^{\text{tp}}(a,y=1,y=1) + \mu^{\text{fp}} f^{\text{fp}}(a,y=0,y=1)} - 0}{\frac{1}{\frac{1}{\mathbb{P}^*(y=1|\mathbf{x},a)} - \mu^{\text{tp}} f^{\text{tp}}(a,y=1,y=1) + \mu^{\text{fp}} f^{\text{fp}}(a,y=0,y=1)} + \frac{1}{\frac{1}{1-\mathbb{P}^*(y=1|\mathbf{x},a)} - 0}} \tag{25}$$

$$\implies Z_{\mathbb{P}(\mathbf{x},a)} = \frac{1 - \mathbb{P}^*(y|\mathbf{x}, a)\mu^{\text{tp}} f^{\text{tp}}(a, y, y)}{1 + (\mu^{\text{fp}} f^{\text{fp}}(a, 1 - y, y) - \mu^{\text{tp}} f^{\text{tp}}(a, y, y))\mathbb{P}^*(y|\mathbf{x}, a) - (\mu^{\text{fp}} f^{\text{fp}}(a, 1 - y, y) - \mu^{\text{tp}} f^{\text{tp}}(a, y, y))\mathbb{P}^{*2}(y|\mathbf{x}, a)} \tag{26}$$

□

For **Equalized Opportunity** where $F^{\text{FP}}(a, ., .) = 0$ we get:

$$Z_{\mathbb{P}(\mathbf{x},a)} = \frac{1 - \mathbb{P}^*(y|\mathbf{x}, a)\mu^{\text{tp}} f^{\text{tp}}(a, y, y)}{1 - \mu^{\text{tp}} f^{\text{tp}}(a, y, y)\mathbb{P}^*(y|\mathbf{x}, a) + \mu^{\text{tp}} f^{\text{tp}}(a, y, y)\mathbb{P}^{*2}(y|\mathbf{x}, a)} \tag{27}$$

$$\implies \mathbb{Q}^*(y|\mathbf{x}, a) = \frac{\mathbb{P}^*(y|\mathbf{x}, a)}{1 - \mu^{\text{tp}} f^{\text{tp}}(a, y, y)\mathbb{P}^*(y|\mathbf{x}, a) + \mu^{\text{tp}} f^{\text{tp}}(a, y, y)\mathbb{P}^{*2}(y|\mathbf{x}, a)} \tag{28}$$

where additionally it must hold that $0 \le \mathbb{Q}^* \le 1$.

$$\implies 0 \le \frac{\mathbb{P}^*(y|\mathbf{x}, a)}{1 - \mu^{\text{tp}} f^{\text{tp}}(a, y, y) + \mu^{\text{tp}} f^{\text{tp}}(a, y, y)\mathbb{P}^{*2}(y|\mathbf{x}, a)} \le 1 \tag{29}$$

$$\overset{y=1}{\implies} \begin{cases} \mathbb{P}(y|\mathbf{x}, a) < \frac{1}{\mu^{\text{tp}} f^{\text{tp}}(a)} & \text{if } \mu^{\text{tp}} f^{\text{tp}}(a) > 1 \\ \mathbb{P}(y|\mathbf{x}, a) < 1 & \text{otherwise} \end{cases} \tag{30}$$