Fair Logistic Regression: An Adversarial Perspective

Ashkan Rezaei [*] [1] **Rizal Fathony** [*] [1] **Omid Memarrast** [1] **Brian D. Ziebart** [1]

Abstract

Fair prediction methods have primarily been built around existing classification techniques using pre-processing methods, post-hoc adjustments, reduction-based constructions, or deep learning procedures. We investigate a new approach to fair data-driven decision making by designing predictors with fairness requirements integrated into their core formulations. We augment a game-theoretic construction of the logistic regression model with fairness constraints, producing a novel prediction model that robustly *and fairly* minimizes the logarithmic loss. We demonstrate the advantages of our approach on a range of benchmark datasets for fairness.

1 Introduction

Though maximizing accuracy has been the principal objective for classification tasks, competing priorities are also often of key concern in practice. Fairness properties guaranteeing different groups are treated equivalently by the classifier in various ways are a prime example. These may be desirable—or even legally required—when making admissions decisions for universities (Lowry & Macpherson, 1988; Chang, 2006; Kabakchieva, 2013), employment and promotion decisions for organizations (Lohr, 2013), medical decisions for hospitals and insurers (Shipp et al., 2002; Obermeyer & Emanuel, 2016), sentencing guidelines within the judicial system (Moses & Chan, 2014; O'Neil, 2016), loan decisions for the financial industry (Shaw & Gentry, 1988; Carter & Catlett, 1987; Bose & Mahapatra, 2001), and in many other applications.

The notion of fairness in decision making fall into two main categories: *group fairness* criteria generally partition the population based on a protected attribute into groups and mandate equal treatment of members across groups based on some defined statistical measures. The notion of *individual fairness* (Dwork et al., 2012) on the other hand, demands that similar individuals should be treated similarly irrespective of group membership.

In this paper we focus on group fairness measures, namely the three prevalent measures of *demographic parity* (Calders et al., 2009), *equalized odds* (Hardt et al., 2016), and *equalized opportunity* (Hardt et al., 2016). Techniques for constructing predictors that provide these fairness guarantees largely leverage existing classification methods as black boxes. Preprocessing methods such as reweighting and relabeling (Kamiran & Calders, 2012) transform the input data to remove dependence between the class and protected attribute according to a predefined fairness constraint. Other preprocessing methods (Calmon et al., 2017; Zemel et al., 2013) cast the transformation as an optimization problem to find a randomized mapping that limits the dependence of the transformed outcome on protected attribute while remaining statistically close to the original dataset. In contrast, post-hoc methods adjust the class labels provided from black box classifiers by mixing them with uninformed predictions to satisfy fairness requirements (Hardt et al., 2016). Though guaranteeing equalized odds and calibration has been shown by Kleinberg et al. (2016) and Chouldechova (2017) to be impossible in general, Pleiss et al. (2017) propose a calibrated postprocessing method that provides a relaxed notion of equalized odds. Zafar et al. (2017) augments any decision boundary classifier with fairness constraints, resulting in a non-convex optimization problem that can be solved using convex-concave programming. Reduction-based fairness methods combine cost-sensitive black box predictors to produce a fair predictor (Agarwal et al., 2018). Generative adversarial techniques have also been employed to construct intermediate feature representations that do not allow group status to be identified (Madras et al., 2018).

In contrast with these previous approaches, we seek to incorporate fairness into the initial construction of a new predictor so that it is designed specifically for the task of fair data-driven decision making. Working from first principles based on robust estimation (Topsøe, 1979; Grünwald & Dawid, 2004; Delage & Ye, 2010), we integrate fairness constraints into the initial formulation of our predictor. We accomplish this by posing the selection of a predictor as an adversarial game between the predictor subjected to fairness constraints on the training sample and an adversarial

[*] Equal contribution [1] Department of Computer Science, University of Illinois at Chicago. Correspondence to: Ashkan Rezaei <arezae4@uic.edu>, Rizal Fathony <rfatho2@uic.edu>.

approximator of the training data labels that must maintain some statistical properties of the training sample.

In the remainder of this paper, we develop our adversarial formulation as a robust log loss minimizer. For some fairness criteria (e.g., equalized opportunity and equalized odds), our approach produces predictive distributions that are conditioned on the actual label. We introduce a fixed point method for making predictions from these conditional distributions and establish the consistency properties of this method. We demonstrate the benefits of our approach compared to existing black-box classification methods on benchmark fair data-driven decision tasks.

2 Background

2.1 Measures of fairness for decision making

Several useful measures have been proposed to quantitatively assess fairness in decision making. Though our approach can be applied to a wider range of fairness constraints, we focus on three prominent ones in this paper: *Demographic Parity* (Calders et al., 2009), *Equality of Opportunity* (Hardt et al., 2016) and *Equality of Odds* (Hardt et al., 2016). We briefly review the variables involved in decision tasks and these three definitions of fairness based on those variables.

For simplicity, we consider a binary decision setting with examples drawn from a distribution: $(\mathbf{X}, A, Y) \sim P$. Here $y = 1$ is viewed as the "advantaged" class for positive decisions to be made. The general decision task is to construct a mapping, $\hat{P}$, for a distribution over decision variable $\hat{y} \in \{0, 1\}$ given the feature vector $\mathbf{x} \in \boldsymbol{\mathcal{X}}$. Each example also possesses a protected attribute $a \in \{0, 1\}$ that defines membership in one of two groups.

We consider three illustrative examples: admissions to medical school, approval of loans, and prescription of medical treatment. The relevant variables are defined for these tasks in Table 1. Different forms of fairness may be appropriate in each of these decision tasks.

Table 1: Variables for three decision-making tasks.

Setting	$\hat{y}$	y	a
Admissions	Admitted	Would succeed	Sex
Loans	Loan approval	Would re-pay	Age
Treatment	Provided	Would benefit	Race

Fairness requires treating the different groups equivalently in various ways. Unfortunately, the naïve approach of excluding the protected attribute from the decision function, e.g., restricting to $\widehat{P}(\hat{y}|\mathbf{x})$, does not guarantee fairness because the protected attribute a may still be inferred from $\mathbf{x}$ (Dwork et al., 2012). For example, graduation from an women's college could serve as an (approximate) surrogate for sex in medical school admissions decisions. Instead of imposing structural constraints on the predictor, various definitions of fairness require properties on its provided decisions to hold.

If student qualifications for medical school admissions are assumed to be the same across sexes, ensuring Demographic Parity (Definition 1) may be the most appropriate form of fairness.

Definition 1. *A classifier satisfies* DEMOGRAPHIC PARITY (D.P.) *if the output variable $\widehat{Y}$ is statistically independent of the protected attribute A:*

$$P(\widehat{Y} = 1|A = a) = P(\widehat{Y} = 1), \quad \forall a \in \mathcal{A}. \qquad (1)$$

If default rates for older (Age ≥ 40) and younger (Age $<$ 40) loan applicants differ, providing the same approval rate to each group (Demographic Parity) may not be desirable. However, providing the same approval rates to individuals who will repay in each group and the same approval rates to individuals who will not repay (Equalized Odds) may be a desirable fairness guarantee.

Definition 2. *A classifier satisfies* EQUALIZED ODDS (E.ODD.) *if the output variable $\widehat{Y}$ is conditionally independent of the protected attribute A given the true label Y:*

$$P(\widehat{Y} = 1|A = a, Y = y) = P(\widehat{Y} = 1|Y = y), \qquad (2)$$
$$\forall y \in \mathcal{Y}, a \in \mathcal{A}.$$

Finally, if there is little benefit from providing a positive decision to a non-advantaged individual, only imposing the above constraint on a particular label (the advantaged class) may be desirable. For example, guaranteeing the same proportion of people who would benefit from a treatment will receive the treatment in each group may be desirable without also requiring the same rates for people who would not benefit from the treatment.

Definition 3. *A classifier satisfies* EQUALIZED OPPORTUNITY (E.OPP.) *if the output variable $\widehat{Y}$ and protected attribute A are conditionally independent given $Y = 1$:*

$$P(\widehat{Y} = 1|A = a, Y = 1) = P(\widehat{Y} = 1|Y = 1), \qquad (3)$$
$$\forall a \in \mathcal{A}.$$

The sets of decision functions $\widehat{P}$ satisfying these fairness constraints are convex and can be defined using linear constraints (Agarwal et al., 2018). The general form for these constraints is:

$$\Gamma : \Big\{\widehat{P} \,|\, \tfrac{1}{p_{\gamma_1}} \mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y) \\ \widehat{P}(\widehat{y}|\mathbf{x},a,y)}} \left[\mathbb{I}(\widehat{Y} = 1 \wedge \gamma_1(A, Y))\right] \qquad (4)$$
$$= \tfrac{1}{p_{\gamma_0}} \mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y) \\ \widehat{P}(\widehat{y}|\mathbf{x},a,y)}} \left[\mathbb{I}(\widehat{Y} = 1 \wedge \gamma_0(A, Y))\right]\Big\},$$

where γ_1 and γ_0 denote some combination of group membership and ground-truth class for each example, while p_{γ_1} and p_{γ_0} denote the empirical frequencies of γ_1 and γ_0: $p_{\gamma_i} = \mathbb{E}_{\widetilde{P}(a,y)}[\gamma_i(A,Y)]$. We can model the fairness constraints (Definitions 1, 2, and 3) as:

$$\Gamma_{\text{dp}} \iff \gamma_j(A,Y) = \mathbb{I}(A=j) \tag{5}$$

$$\Gamma_{\text{e.opp}} \iff \gamma_j(A,Y) = \mathbb{I}(A=j \wedge Y=1) \tag{6}$$

$$\Gamma_{\text{e.odd}} \iff \gamma_j(A,Y) = \begin{bmatrix} \mathbb{I}(A=j \wedge Y=1) \\ \mathbb{I}(A=j \wedge Y=0) \end{bmatrix}. \tag{7}$$

2.2 Robust log-loss minimization, maximum entropy, and logistic regression

The logarithmic loss, $-\sum_{\mathbf{x},y} P(\mathbf{x},y)\log\widehat{P}(y|\mathbf{x})$, is an information-theoretic measure of the expected amount of "surprise" (in bits if using $\log_2$) that the predictor, $\widehat{P}(y|\mathbf{x})$, experiences when encountering labels y distributed according to $P(\mathbf{x},y)$. Robust minimization of the logarithmic loss serves a fundamental role in constructing probability distributions (e.g., Gaussian, Laplacian, Beta, Gamma, and Bernoulli (Lisman & Zuylen, 1972)) and predictors (Manning & Klein, 2003). For conditional probabilities, it is equivalent to maximizing the conditional entropy (Jaynes, 1957):

$$\min_{\widehat{P}(\widehat{y})|\mathbf{x})\in\Delta}\ \max_{\check{P}(\check{y}|\mathbf{x})\in\Delta\cap\Xi} -\sum_{\mathbf{x},y}\check{P}(\mathbf{x},y)\log\widehat{P}(y|\mathbf{x}) \tag{8}$$

$$= \max_{\widehat{P}(\widehat{y}|\mathbf{x})\in\Xi} -\sum_{\mathbf{x},y}\widehat{P}(\mathbf{x},y)\log\widehat{P}(y|\mathbf{x}) = \max_{\widehat{P}(\widehat{y}|\mathbf{x})\in\Xi} H(Y|\mathbf{X}),$$

after simplifications based on the fact that the saddlepoint solution is $\widehat{P} = \check{P}$. When the adversarial distribution is constrained to match the statistics of training data,

$$\Xi : \left\{ \check{P} \mid \mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y) \\ \check{P}(\widehat{y}|\mathbf{x},a,y)}} [\phi(\mathbf{X},\widehat{Y})] = \mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)}[\phi(\mathbf{X},Y)] \right\}, \tag{9}$$

the robust log loss minimizer/maximum entropy predictor (Eq. (8)) is the logistic regression model, $P(y|\mathbf{x}) \propto e^{\theta^{\mathrm{T}}\phi(\mathbf{x},y)}$, with θ estimated by maximizing data likelihood (Manning & Klein, 2003). While technically this distribution needs to only be defined at input values in which training data exists (i.e., $\tilde{P}(\mathbf{x}) > 0$), an inductive assumption that generalizes the form of the distribution to other inputs is employed.

The flexibility of this formulation has been leveraged to provide robust prediction methods for addressing covariate shift (i.e., differing training and testing input distributions) by defining expected log loss and feature-matching constraints under two different covariate distributions (Liu & Ziebart, 2014), and for constructing consistent surrogate losses for general multiclass classification by defining robust predictors under the multiclass loss metrics (Asif et al., 2015; Fathony et al., 2016; 2017; 2018).

Our approach similarly extends this fundamental formulation by imposing fairness constraints on $\widehat{P}$. Since $\Gamma \not\subseteq \Xi$, the saddle point solution does not reflect equality ($\widehat{P} \neq \check{P}$) and therefore is much more complex.

3 Formulation & Algorithms

Given fairness requirements for a predictor (Eq. (4)) and partial knowledge of the population distribution provided by a training sample (Eq. (9)), how should a fair predictor be constructed? Like all inductive reasoning, good performance on a known training sample does not ensure good performance on the unknown population distribution. We take an adversarial perspective in this paper by seeking the best solution for the worst-case population distribution under these constraints.

3.1 Robust and fair log loss minimization

We begin by formulating selection of the robust fair predictor as a minimax game between a fair predictor and an adversarial approximator of the population distribution. We assume the availability of a set of training samples, $\{(\mathbf{x}_i, a_i, y_i)\}_{i=1:n}$, which we equivalently denote by probability distribution $\widetilde{P}(\mathbf{x},a,y)$.

Definition 4. *The* **Fair Robust Log-Loss Predictor**, $\widehat{P}$, *minimizes the worst-case log loss—as chosen by adversary* $\check{P}$ *constrained to reflect training statistics—while providing empirical fairness guarantees:*

$$\min_{\widehat{P}\in\Delta\cap\Gamma}\ \max_{\check{P}\in\Delta\cap\Xi} \mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y) \\ \check{P}(\widehat{y}|\mathbf{x},a,y)}} \left[-\log\widehat{P}(\widehat{Y}|\mathbf{X},A,Y)\right], \tag{10}$$

where Γ *(Eq.* (4)*) and* Ξ *(Eq.* (9)*) are sets enforcing fairness and training data-matching constraints, respectively, and* Δ *is the set of conditional probability simplex constraints (i.e.,* $P(y|\mathbf{x},a) \geq 0, \forall\mathbf{x},y,a; \sum_y P(y|\mathbf{x},a) = 1, \forall\mathbf{x},a$*).*

Though conditioning the decision variable $\widehat{Y}$ on the true label Y would appear to introduce a trivial solution (i.e., $\widehat{Y} = Y$), instead Y only influences $\widehat{Y}$ directly based on fairness properties due to the adversarial construction the predictor. Note that if fairness constraints are removed, then the influence of Y on $\hat{Y}$ disappears (i.e., $\widehat{P}(\widehat{Y}|\mathbf{X},A,Y=0) = \widehat{P}(\widehat{Y}|\mathbf{X},A,Y=1)$) and this formulation ultimately reduces to the familiar logistic regression model (Manning & Klein, 2003) described in §2.2.

This saddlepoint problem has a key beneficial characteristic for optimization: it is convex-concave in $\widehat{P}$ and $\check{P}$ with additional convex constraints (Γ and Ξ) on each distribution.

3.2 Parametric Distribution Form

By leveraging strong minimax duality (Von Neumann & Morgenstern, 1945; Sion, 1958), we derive the parametric form of our predictor as stated in Theorem 1[1].

Theorem 1. *The* **Fair Robust Log-Loss Predictor** *(Definition 4) can be equivalently solved in its dual formulation:*

$$\begin{aligned}\min_{\theta}\max_{\lambda}\ & \frac{1}{n}\sum_{(\mathbf{x},a,y)\in\mathcal{D}}\Big\{\mathbb{E}_{\check{P}_{\theta,\lambda}(\widehat{y}|\mathbf{x},a,y)}\Big[-\log\widehat{P}_{\theta,\lambda}(\widehat{Y}|\mathbf{x},a,y)\Big] \\ & + \theta^{\top}\Big(\mathbb{E}_{\check{P}_{\theta,\lambda}(\widehat{y}|\mathbf{x},a,y)}[\phi(\mathbf{x},\widehat{Y})] - \phi(\mathbf{x},y)\Big) \\ & + \lambda\big(\tfrac{1}{p_{\gamma_1}}\mathbb{E}_{\widehat{P}_{\theta,\lambda}(\widehat{Y}|\mathbf{x},a,y)}[\mathbb{I}(\widehat{Y}=1\wedge\gamma_1(A,Y))] \\ & \quad - \tfrac{1}{p_{\gamma_0}}\mathbb{E}_{\widehat{P}_{\theta,\lambda}(\widehat{Y}|\mathbf{x},a,y)}[\mathbb{I}(\widehat{Y}=1\wedge\gamma_0(A,Y))]\big)\Big\},\end{aligned} \tag{11}$$

where θ and λ are Lagrange multipliers for the moment matching and fairness constraints, respectively, and n is the number of samples in the dataset. The parametric distribution of $\widehat{P}$ is defined when $\lambda > 0$, as:

$$\widehat{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y) = \tag{12}$$
$$\begin{cases}\min\left\{\frac{\exp(\theta^{\top}\phi(\mathbf{x},1))}{Z_\theta(\mathbf{x})}, \frac{p_{\gamma_1}}{\lambda}\right\} & \textit{if } \gamma_1(a,y) \\ \max\left\{\frac{\exp(\theta^{\top}\phi(\mathbf{x},1))}{Z_\theta(\mathbf{x})}, 1-\frac{p_{\gamma_0}}{\lambda}\right\} & \textit{if } \gamma_0(a,y) \\ \frac{\exp(\theta^{\top}\phi(\mathbf{x},1))}{Z_\theta(\mathbf{x})} & \textit{otherwise};\end{cases}$$

when $\lambda < 0$, as:

$$\widehat{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y) = \tag{13}$$
$$\begin{cases}\max\left\{\frac{\exp(\theta^{\top}\phi(\mathbf{x},1))}{Z_\theta(\mathbf{x})}, 1+\frac{p_{\gamma_1}}{\lambda}\right\} & \textit{if } \gamma_1(a,y) \\ \min\left\{\frac{\exp(\theta^{\top}\phi(\mathbf{x},1))}{Z_\theta(\mathbf{x})}, -\frac{p_{\gamma_0}}{\lambda}\right\} & \textit{if } \gamma_0(a,y) \\ \frac{\exp(\theta^{\top}\phi(\mathbf{x},1))}{Z_\theta(\mathbf{x})} & \textit{otherwise};\end{cases}$$

and when $\lambda = 0$, as:

$$\widehat{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y) = \frac{\exp(\theta^{\top}\phi(\mathbf{x},1))}{Z_\theta(\mathbf{x})}, \tag{14}$$

where $Z_\theta(\mathbf{x}) = \exp(\theta^{\top}\phi(\mathbf{x},1)) + \exp(\theta^{\top}\phi(\mathbf{x},0))$ is the normalization constant. The parametric distribution of $\check{P}$ is defined using the following relationship with $\widehat{P}$:

$$\check{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y) = \widehat{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y)\times \tag{15}$$
$$\begin{cases}\big(1+\frac{\lambda}{p_{\gamma_1}}\widehat{P}_{\theta,\lambda}(\widehat{y}=0|\mathbf{x},a,y)\big) & \textit{if } \gamma_1(a,y) \\ \big(1-\frac{\lambda}{p_{\gamma_0}}\widehat{P}_{\theta,\lambda}(\widehat{y}=0|\mathbf{x},a,y)\big) & \textit{if } \gamma_0(a,y) \\ 1 & \textit{otherwise}.\end{cases}$$

Note that the predictor's parametric distribution is similar to the parametric distribution of standard binary logistic regression, with the option to *truncate* the probability based on the value of λ. The truncation of $\widehat{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y)$ is from above when $0 < \frac{p_{\gamma_1}}{\lambda} < 1$ and $\gamma_1(a,y) = 1$, and from below when $-1 < \frac{p_{\gamma_1}}{\lambda} < 0$ and $\gamma_1(a,y) = 1$. The adversary's parametric distribution can be computed from the predictor's distribution using the quadratic function in Eq. (15), for example in the case where $\gamma_1(a,y) = 1$:

$$\begin{aligned}\check{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y) =& \rho(1+\tfrac{\lambda}{p_{\gamma_1}}(1-\rho)) \\ =& (1+\tfrac{\lambda}{p_{\gamma_1}})\rho - \tfrac{\lambda}{p_{\gamma_1}}\rho^2,\end{aligned} \tag{16}$$

where $\rho = \widehat{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y)$.

Figure 1 describes the relation between $\widehat{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y)$ and $\check{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y)$ in the case of $\gamma_1(a,y) = 1$. When $\frac{\lambda}{p_{\gamma_1}} = 0$, the adversary's probability is equal to the predictor's probability as shown in the plot as a straight line. Positive values of λ shift the line upward and change it into a curve (e.g., $\frac{\lambda}{p_{\gamma_1}} = 1$) as shown in the plot. As the value of λ grows larger (e.g., $\frac{\lambda}{p_{\gamma_1}} = 2$), some of the valid predictor probabilities ($0 < \widehat{P} < 1$) map to invalid adversary probabilities (i.e., $\check{P} \geq 1$) according to the quadratic function. This occurs for the case of $\frac{\lambda}{p_{\gamma_1}} = 2$ when $\widehat{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y) > 0.5$. In this case, the predictor's probability is truncated according to Eq. (12) such that it cannot exceed $\frac{p_{\gamma_1}}{\lambda} = 0.5$. Similarly, when λ is negative, the curve is shifted downward and the predictor's probability is truncated when the quadratic function mapping results in a negative value of $\check{P}$. For the case where $\gamma_0(a,y) = 1$, the reverse shifting is observed, i.e., shifting downward when λ is positive and shifting upward when λ is negative.

3.3 Enforcing fairness constraints

The inner maximization in Eq. (11) aims to find the optimal λ that enforces the fairness constraint. From the perspective of the parametric distribution of $\widehat{P}$, this is equivalent with finding the threshold points (e.g., $\frac{p_{\gamma_1}}{\lambda}$ and $1 - \frac{p_{\gamma_0}}{\lambda}$) in the $\min$ and $\max$ function of the Eq. (12) such that the expectation of the truncated exponential probabilities of $\widehat{P}$ in group γ_1 match the one in group γ_0. Given the value of θ, we find the optimum λ^* directly by finding the threshold. We first compute the exponential probabilities $P_e(\widehat{y}=1|\mathbf{x},a,y) = \exp(\theta^{\top}\phi(\mathbf{x},1))/Z_\theta(\mathbf{x})$ for each examples in γ_1 and γ_0. Let E_1 and E_0 be the sets that contains P_e for group γ_1 and γ_0 respectively, and let $\bar{e}_1$ and $\bar{e}_0$ be the average of E_1 and E_0 respectively.

Finding λ^* given the sets E_1 and E_0 requires sorting the probabilities for each set, and then iteratively finding the threshold points for both sets (t_1 and t_0 respectively) si-

[1] The proofs of this theorem and other theorems in the paper are available in the supplementary material.

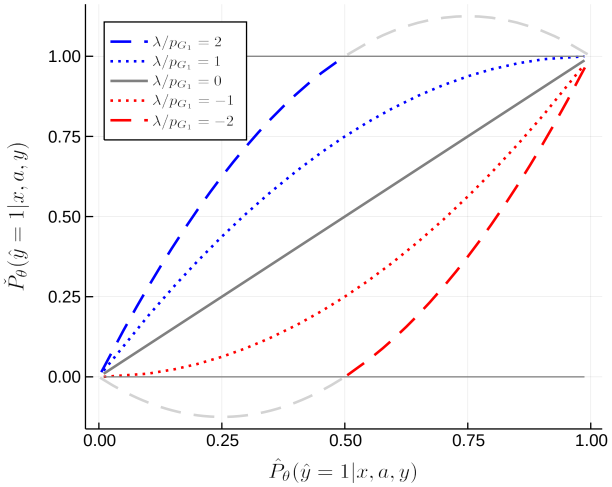

Figure 1: Plot of the relation between predictor and adversary's parametric distributions.

multaneously as described in Algorithm 1. Without loss of generality[2], the algorithm assumes that $\bar{e}_1 > \bar{e}_0$. The algorithm proceeds by first setting the threshold for E_1 as 1, and the threshold for E_0 as 0, i.e., no truncation. It then moves the threshold towards the next probability in the list while maintaining the probability gain (i.e., the difference between the exponential probability and the truncated probability) if it moves the threshold. The algorithm stop when the probability gain is equal to the difference between $\bar{e}_1$ and $\bar{e}_0$.

The runtime of Algorithm 1 is dominated by sorting, i.e., $O(n \log n)$ time. However, if we perform gradient based optimization on θ, the value of the current θ in each iteration does not change much, and neither do the exponential probabilities in each group. By maintaining the sorted list in each iteration as the basis index, the next iteration will have the probabilities in a nearly sorted order. Therefore, we can improve the sorting cost requirement by running sorting algorithms that work best on a nearly sorted list (e.g. insertion sort, Timsort, or P^3-sort) with run times approaching $O(n)$ as the list is closer to a fully sorted list (Chandramouli & Goldstein, 2014).

3.4 Learning

Our learning process seeks the parameters θ, λ for our distributions ($\widehat{P}_{\theta,\lambda}$ and $\check{P}_{\theta,\lambda}$) that match the statistics of the adversary's distribution with training data (θ) and seek fairness (λ), as illustrated in Eq. (11). These parametric distributions are then used to make predictions for the dataset

Algorithm 1 Find λ^* given E_1 and E_0

```
Input: $(E_1, E_0)$, s.t. $\bar{e}_1 > \bar{e}_0$
Sort $E_1$ in decreasing order
Sort $E_0$ in increasing order
Calculate the difference $\bar{d} = \bar{e}_1 - \bar{e}_0$
$t_1 \leftarrow 1,\ t_0 \leftarrow 0$ {thresholds for $E_1$ and $E_0$ resp.}
Set gain to be 0.
while the gain is less than $\bar{d}$ do
    Calculate two candidate for the next move:
      (1)  move $t_1$ to the next $P_e$ in $E_1$ list
      (2)  move $t_0$ to the next $P_e$ in $E_0$ list
    Calculate the gain for each move and the effect of
    the move for the other group.
    Choose the move that has the minimum gain
end while
Calculate the threshold that produces gain equal to $\bar{d}$,
which is located between the last move in the loop and
the threshold before the move
Calculate $\lambda^*$ based on the threshold
Return: $\lambda^*$
```

that the learner has not seen with the hope that the knowledge the learner obtains from the training dataset can be transferred to this test dataset.

In the previous subsection, we derived an algorithm to directly compute the best λ given arbitrary value of θ. Let λ^*_θ be this optimal solution of the inner optimization in Eq. (11). Given λ^*_θ, the optimization of Eq. (11) reduces into a simpler optimization solely over θ, as described in Theorem 2.

Theorem 2. *Given the optimum value of λ^*_θ for θ, the dual formulation in Eq.* (11) *reduces to the minimization over:*

$$\mathcal{L}(\mathcal{D}) = \min_\theta \frac{1}{n} \sum_{(\mathbf{x},a,y)\in\mathcal{D}} \ell_{\theta,\lambda^*_\theta}(\mathbf{x}, a, y), \tag{17}$$

where, for the case that $\lambda^*_\theta > 0$, ℓ is defined as:

$$\ell_{\theta,\lambda^*}(\mathbf{x}, a, y) = -\theta^\top \phi(\mathbf{x}, y) + \tag{18}$$

$$\begin{cases} -\log(\frac{p_{\gamma_1}}{\lambda^*_\theta}) + \theta^\top(\phi(\mathbf{x}, 1)) & \textit{if } \gamma_1(a, y) \wedge T(\mathbf{x}, \theta) \\ -\log(\frac{p_{\gamma_0}}{\lambda^*_\theta}) + \theta^\top(\phi(\mathbf{x}, 0)) & \textit{if } \gamma_0(a, y) \wedge T(\mathbf{x}, \theta) \\ \log(Z_\theta(\mathbf{x})) & \textit{otherwise}, \end{cases}$$

for the case that $\lambda^*_\theta < 0$, ℓ is defined as:

$$\ell_{\theta,\lambda^*}(\mathbf{x}, a, y) = -\theta^\top \phi(\mathbf{x}, y) + \tag{19}$$

$$\begin{cases} -\log(-\frac{p_{\gamma_1}}{\lambda^*_\theta}) + \theta^\top(\phi(\mathbf{x}, 0)) & \textit{if } \gamma_1(a, y) \wedge T(\mathbf{x}, \theta) \\ -\log(-\frac{p_{\gamma_0}}{\lambda^*_\theta}) + \theta^\top(\phi(\mathbf{x}, 1)) & \textit{if } \gamma_0(a, y) \wedge T(\mathbf{x}, \theta) \\ \log(Z_\theta(\mathbf{x})) & \textit{otherwise}, \end{cases}$$

and for the case that $\lambda^*_\theta = 0$, ℓ is defined as:

$$\ell_{\theta,\lambda^*_\theta}(\mathbf{x}, a, y) = \log(Z_\theta(\mathbf{x})) - \theta^\top \phi(\mathbf{x}, y). \tag{20}$$

[2] For the case when $\bar{e}_1 < \bar{e}_0$, we flip the group membership and then λ^* is the negative of the solution produced by the algorithm. In the case of $\bar{e}_1 = \bar{e}_0$ the exponential probabilities are already fair, and we set $\lambda^* = 0$.

Here, $T(\mathbf{x},\theta)$ returns 1 if the exponential probability is truncated (for example when $\frac{\exp(\theta^\top \phi(x,1))}{Z_\theta(x)} > \frac{p_{\gamma_1}}{\lambda^*_\theta}$ in the case where $\gamma_1(a,y) = 1$ and $\lambda^*_\theta > 0$), or 0 otherwise.

To solve for θ, we use a gradient-based optimization technique. From the objective in Eq. (17), we derive the gradient of the objective with respect to θ in Theorem 3.

Theorem 3. *The (sub)-gradient of the learning objective (Eq.* (17)*) with respect to θ contains:*

$$\frac{1}{n} \sum_{(\mathbf{x},a,y)\in\mathcal{D}} g_{\theta,\lambda^*_\theta}(\mathbf{x},a,y) \in \partial_\theta \mathcal{L}, \textit{ where:} \tag{21}$$

(1) $g_{\theta,\lambda^*_\theta}(\mathbf{x},a,y) = \phi(x,1) - \phi(\mathbf{x},y)$

if $(\gamma_1(a,y) \wedge T(\mathbf{x},\theta) \wedge \mathbb{I}[\lambda^*_\theta > 0])$

$\vee\ (\gamma_0(a,y) \wedge T(\mathbf{x},\theta) \wedge \mathbb{I}[\lambda^*_\theta < 0])$

(2) $g_{\theta,\lambda^*_\theta}(\mathbf{x},a,y) = \phi(x,0) - \phi(\mathbf{x},y)$

if $(\gamma_0(a,y) \wedge T(\mathbf{x},\theta) \wedge \mathbb{I}[\lambda^*_\theta > 0])$

$\vee\ (\gamma_1(a,y) \wedge T(\mathbf{x},\theta) \wedge \mathbb{I}[\lambda^*_\theta < 0])$

(3) $g_{\theta,\lambda^*_\theta}(\mathbf{x},a,y) = \sum_{y'\in\mathcal{Y}} \frac{\exp(\theta^\top \phi(\mathbf{x},y'))}{Z_\theta(x)} \phi(\mathbf{x},y') - \phi(\mathbf{x},y)$

otherwise.

To improve the generalizability of our parametric model, we employ a standard L2 regularization technique, which is also used in many machine learning algorithms including the standard logistic regression model, i.e., we optimize for:

$$\theta^* = \operatorname*{argmin}_\theta \frac{1}{n} \sum_{(\mathbf{x},a,y)\in\mathcal{D}} \ell_{\theta,\lambda^*_\theta}(\mathbf{x},a,y) + \tfrac{C}{2}\|\theta\|_2^2, \tag{22}$$

where C is the regularization constant. We employ a standard batch gradient descent optimization algorithm (e.g., L-BFGS) to obtain the optimal θ^* of Eq. (22). We also compute the corresponding optimal inner optimization, $\lambda^*_{\theta^*}$. We then construct the optimal predictor and adversary's parametric distributions based on the value of θ^* and $\lambda^*_{\theta^*}$.

3.5 Inference

In the inference step, we apply the optimum predictor parametric distribution $\widehat{P}_{\theta^*,\lambda^*_{\theta^*}}$ we learn from the training step to the new examples in the testing set. Given the value of θ^* and $\lambda^*_{\theta^*}$, we calculate the predictor's distribution for our new data point using Eq. (12), (13), or (14) depending on the value $\lambda^*_{\theta^*}$. Note that the predictor's parametric distribution also depends on the group membership of the example. For fairness constraints that solely based on the protected attribute A, e.g., D.P., these parametric formulation can be directly applied. However, some of the fairness constraints also depend on the true label, e.g., E.OPP. and E.ODD.. In this cases, we construct a prediction procedure that approximate the true label with the adversary's parametric distribution.

For the fairness constraints that depend on the true label, our algorithm output the predictor and adversary's parametric distributions condition on the value of true label, i.e., $\widehat{P}(\widehat{y}|\mathbf{x},a,y=1)$, $\widehat{P}(\widehat{y}|\mathbf{x},a,y=0)$, $\widecheck{P}(\widehat{y}|\mathbf{x},a,y=1)$, and $\widecheck{P}(\widehat{y}|\mathbf{x},a,y=0)$. Our goal is to produce conditional probability of $\widehat{y}$ that does not depends on the true label, i.e., $\widehat{P}(\widehat{y}|\mathbf{x},a)$. We construct the following procedure to estimate this probability.

Based on the marginal probability rule, $\widehat{P}(\widehat{y}|\mathbf{x},a)$ can be expressed as follows:

$$\begin{aligned}\widehat{P}(\widehat{y}|\mathbf{x},a) = {} & \widehat{P}(\widehat{y}|\mathbf{x},a,y=1)P(y=1|\mathbf{x},a) \\ & + \widehat{P}(\widehat{y}|\mathbf{x},a,y=0)P(y=0|\mathbf{x},a).\end{aligned} \tag{23}$$

However, since we do not have access to $P(y|x,a)$, we cannot directly apply the formulation. Instead, we approximate $P(y|x,a)$ with the adversary's distribution $\widecheck{P}(\widehat{y}|x,a)$. Using the similar marginal probability rule, we express the estimate as follows:

$$\begin{aligned}\widecheck{P}(\widehat{y}|\mathbf{x},a) \approx {} & \widecheck{P}(\widehat{y}|\mathbf{x},a,y=1)\widecheck{P}(\widehat{y}=1|\mathbf{x},a) \\ & + \widecheck{P}(\widehat{y}|\mathbf{x},a,y=0)\widecheck{P}(\widehat{y}=0|\mathbf{x},a).\end{aligned} \tag{24}$$

By rearranging the terms in the formulation above, we calculate the approximation as:

$$\widecheck{P}(\widehat{y}=1|\mathbf{x},a) = \frac{\widecheck{P}(\widehat{y}=1|\mathbf{x},a,y=0)}{\widecheck{P}(\widehat{y}=0|\mathbf{x},a,y=1)+\widecheck{P}(\widehat{y}=1|\mathbf{x},a,y=0)}, \tag{25}$$

which can be directly computed from the adversary's parametric distribution produced by our model using Eq. (15). Finally, to get the approximation over the predictor's conditional probability ($\widehat{P}(\widehat{y}|\mathbf{x},a)$), we replace $P(y|\mathbf{x},a)$ in Eq. (23) with $\widecheck{P}(\widehat{y}|\mathbf{x},a)$ calculated from Eq. (25).

3.6 Asymptotic convergence property

The behavior of an algorithm in the limit should be considered when designing a learning algorithm. Indeed, the asymptotic convergence property studies the learning setting when a learning algorithm is provided with access to the true data generating distribution $P(\mathbf{x},a,y)$ and a fully expressive feature representation. We show in Theorem 4 that in the limit, our method finds a predictor distribution that has a desirable characteristic in terms of the Kullback-Leibler (KL) divergence from the true distribution.

Theorem 4. *Given access to the true population distribution $P(\mathbf{x},a,y)$ and a fully expressive feature representation, our formulation (Definition 4) finds the predictor* *with the closest distance to $P(\mathbf{x}, a, y)$ in terms of the KL-divergence, from the set of probability distributions that satisfy the fairness constraints.*

We next show in Theorem 5 that for the case where the fairness constraint depends on the true label (e.g., E.OPP. and E.ODD.), our prediction procedure in §3.5 outputs a predictor distribution with the same desired characteristic, marginalized over the true label.

Theorem 5. *For fairness constraints that depend on the true label, our inference procedure in §3.5 produces the marginal predicting distribution of the fair predictor distribution with the closest KL-divergence distance to $P(\mathbf{x}, a, y)$ in the limit.*

4 Experiments

4.1 Datasets

We evaluate our proposed algorithm on three benchmark datasets on fairness:

- The UCI Adult (Dheeru & Karra Taniskidou, 2017) dataset includes 45,222 samples with an income greater than $50k considered to be a favorable binary outcome. We choose gender as the protected attribute. 11 other features (age, race, workclass, education-num, marital-status, occupation, relationship, capital-gain, capital-loss, hours-per-week, native-country) are available for each example to use for basing decisions.
- The ProPublica's COMPAS recidivism dataset (Larson et al., 2016) contains 6,167 samples, and the task is to predict the recidivism of an individual based on criminal history, with the binary protected attribute being race (white and non-white). An additional nine features (sex, age, age_cat, juv_fel_count, juv_misd_count, juv_other_count, priors_count, c_charge_degree, c_charge_desc) are available for basing decisions.
- The dataset from the Law School Admissions Councils National Longitudinal Bar Passage Study (Wightman, 1998) has 20,649 examples. Here the favorable outcome for the individual is passing the bar exam, and race (restricted to white and black only) is the protected attribute. 13 other features available for basing decisions.

4.2 Comparison methods

We compare our method against various learning algorithms designed to provide fair predictions:

- The **unconstrained logistic regression** model (*LR_unconstrained*) is a standard logistic regression model that ignores all fairness requirements.
- The **cost sensitive reduction approach** (*reduction*) by Agarwal et al. (2018) reduces fair classification to learning a randomized hypothesis over a sequence of cost-sensitive classifiers. We use the sample-weighted implementation of Logistic Regression in scikit-learn (Pedregosa et al., 2011) as the base classifier, to compare the effect of reduction approach. We evaluate the performance of the model by varying the constraint bounds across the set $\epsilon \in \{.001, .01, .1\}$.
- For demographic parity, we compare with the **reweighting method** (*reweight*) of Kamiran & Calders (2012), which learns weights for each combination of class label and protected attribute and then uses these weights to resample from the original training data which yields a new dataset with no statistical dependence between class label and protected attribute. The new balanced dataset is then used for training a classifier. We use the implementation of this approach in IBM toolkit (Bellamy et al., 2018).
- For equalized odds, we also compare with the **post-processing method** (*postprocessing*) of Hardt et al. (2016) which transforms the classifier's output by solving a linear program that find a probabilistic prediction which minimizes the misclassification error and satisfies the equalized odds constraint from the set of probability formed by the convex hull of the original classifier's probabilities and the extreme point of probability values (i.e, zero and one).

4.3 Evaluation measures and setup

Data-driven fair decision methods seek to minimize both prediction error rates and measures of unfairness. We consider the misclassification rate (i.e., the 0-1 loss, $\mathbb{E}[\hat{Y} \neq Y]$) on a withheld test sample to measure prediction error. To quantify the unfairness of each method, we measure the degree of fairness violation for demographic parity (D.P.) as:

$$\left|\mathbb{E}\left[\mathbb{I}(\hat{Y}=1)\middle|A=1\right]-\mathbb{E}\left[\mathbb{I}(\hat{Y}=1)\middle|A=0\right]\right|, \quad (26)$$

and the sum of fairness violations for each class to measure the total violation for equalized odds (E.ODD.):

$$\sum_{y\in\{0,1\}}\left(\left|\mathbb{E}\left[\mathbb{I}(\widehat{Y}=1)\middle|A=1,Y=y\right]-\right.\right. \quad (27)$$
$$\left.\left.\mathbb{E}\left[\mathbb{I}(\widehat{Y}=1)\middle|A=0,Y=y\right]\right|\right),$$

to obtain a level comparison across different methods.

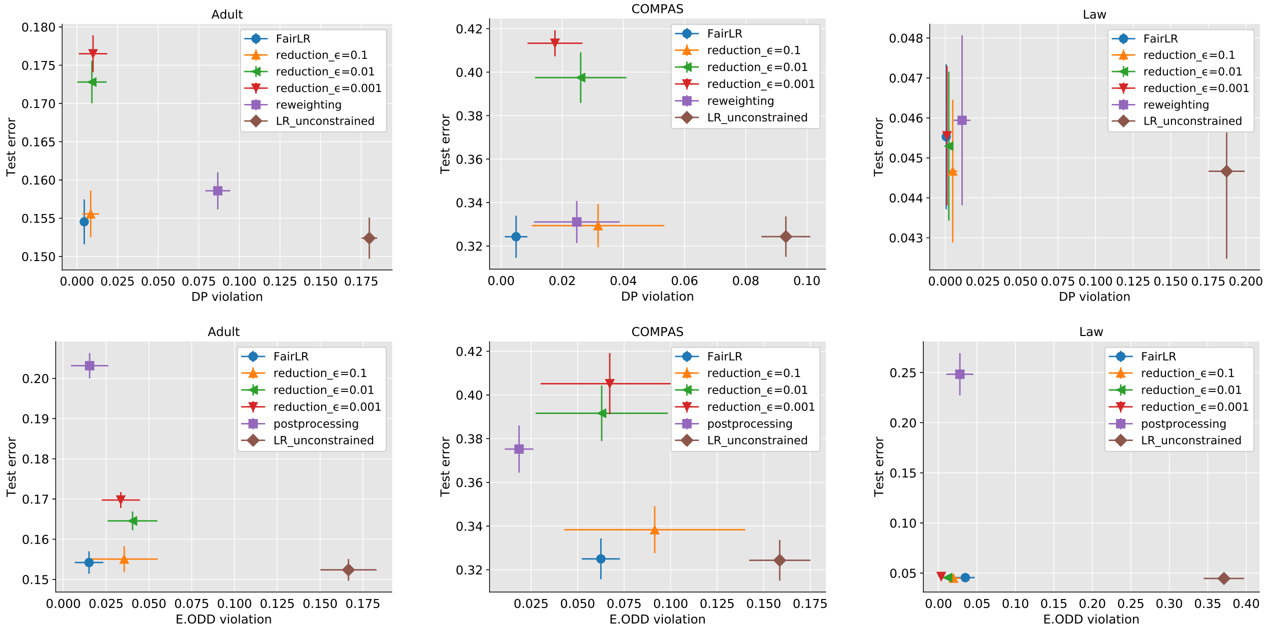

Figure 2: *Test classification error* versus *Demographic Parity* constraint violation (top row) and *Equalized Odd* constraint violation (bottom row). The bars indicate standard deviation on 20 random splits of data.

We perform all of our experiments using 20 random splits of each dataset into a training set (70% of examples) and a testing set (30%). We record the averages over these twenty random splits and the standard deviation. We cross validate our model on a separate validation set using the best logloss to select an L2 penalty from $(\{.001, .005, .01, .05, .1, .2, .3, .4, .5\})$.

4.4 Experimental Results

Figure 2 provides the evaluation results (test error and fairness violation) of each method for demographic parity and equalized odds on test data from each of the three datasets (for training performance, see Appendix B). Fairness can be vacuously achieved by an agnostic predictor that always outputs 1 or always outputs 0. Thus, the appropriate question to ask when considering these results is: "how much additional test error is incurred compared to the baseline of the unfair predictor (*LR_unconstrained*) for how much of an increase in fairness?"

For demographic parity on the Adult and COMPAS datasets, our FairLR approach outperforms all baseline methods seeking fairness on average for both test error rate and for fairness violations. Additionally, the increase in test error over the unfair unconstrained logistic regression model is small. For demographic parity on the Law dataset, the relationship between methods is not as clear, but our FairLR approach still resides in the Pareto optimal set, i.e., there are no other methods that are better than our result on both criteria.

For equalized odds, FairLR provides the lowest ratios for increasing fairness over increasing error rate for the Adult and COMPAS datasets, and competitive performance on the Law dataset. The post-processing method provides comparable or better fairness at the cost of significantly higher error rates.

5 Conclusions & Future Work

We have developed a novel approach for providing fair data-driven decision making in this work by deriving a modified logistic regression model from first principles. We used a robust estimation formulation (Topsøe, 1979; Grünwald & Dawid, 2004; Delage & Ye, 2010) that imposes fairness requirements on the predictor and views uncertainty about the population distribution pessimistically while maintaining a semblance of the training data characteristics through feature-matching constraints.

Though we focus on classification tasks, our formulation is quite general due to the flexibility of its adversarial construction. In future work, we plan to extend this approach to fairness to multivariate prediction tasks, including learning to fairly rank (Singh & Joachims, 2018) and learning to provide fair assignments. Motivated by recent work on dynamic fairness as a property of processes (Hashimoto et al., 2018), we also plan to investigate the application of fairness to covariate shift settings (Liu et al., 2015).

References

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

Asif, K., Xing, W., Behpour, S., and Ziebart, B. D. Adversarial cost-sensitive classification. In *UAI*, pp. 92–101, 2015.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018. URL `https://arxiv.org/abs/1810.01943`.

Bose, I. and Mahapatra, R. K. Business data mininga machine learning perspective. *Information & management*, 39(3):211–225, 2001.

Calders, T., Kamiran, F., and Pechenizkiy, M. Building classifiers with independency constraints. In *Data mining workshops, 2009. ICDMW'09. IEEE international conference on*, pp. 13–18. IEEE, 2009.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems 30*, pp. 3992–4001. 2017.

Carter, C. and Catlett, J. Assessing credit card applications using machine learning. *IEEE expert*, 2(3):71–79, 1987.

Chandramouli, B. and Goldstein, J. Patience is a virtue: Revisiting merge and sort on modern processors. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pp. 731–742. ACM, 2014.

Chang, L. Applying data mining to predict college admissions yield: A case study. *New Directions for Institutional Research*, 2006(131):53–68, 2006.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *CoRR*, abs/1703.00056, 2017. URL `http://arxiv.org/abs/1703.00056`.

Delage, E. and Ye, Y. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

Dheeru, D. and Karra Taniskidou, E. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226. ACM, 2012.

Fathony, R., Liu, A., Asif, K., and Ziebart, B. Adversarial multiclass classification: A risk minimization perspective. In *Advances in Neural Information Processing Systems 29*, pp. 559–567. 2016.

Fathony, R., Bashiri, M. A., and Ziebart, B. Adversarial surrogate losses for ordinal regression. In *Advances in Neural Information Processing Systems*, pp. 563–573. 2017.

Fathony, R., Asif, K., Liu, A., Bashiri, M. A., Xing, W., Behpour, S., Zhang, X., and Ziebart, B. D. Consistent robust adversarial prediction for general multiclass classification. *arXiv preprint arXiv:1812.07526*, 2018.

Grünwald, P. D. and Dawid, A. P. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, 32:1367–1433, 2004.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

Hashimoto, T. B., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. *arXiv preprint arXiv:1806.08010*, 2018.

Jaynes, E. T. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Kabakchieva, D. Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies*, 13(1):61–72, 2013.

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

Kleinberg, J. M., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *CoRR*, abs/1609.05807, 2016. URL `http://arxiv.org/abs/1609.05807`.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9, 2016.

Lisman, J. and Zuylen, M. v. Note on the generation of most probable frequency distributions. *Statistica Neerlandica*, 26(1):19–23, 1972.

Liu, A. and Ziebart, B. Robust classification under sample selection bias. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 37–45. 2014.

Liu, A., Reyzin, L., and Ziebart, B. D. Shift-pessimistic active learning using robust bias-aware prediction. In *AAAI*, pp. 2764–2770, 2015.

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

Lowry, S. and Macpherson, G. A blot on the profession. *British medical journal (Clinical research ed.)*, 296 (6623):657, 1988.

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

Manning, C. and Klein, D. Optimization, maxent models, and conditional estimation without magic. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Tutorials-Volume 5*, pp. 8. Association for Computational Linguistics, 2003.

Moses, L. B. and Chan, J. Using big data for legal and law enforcement decisions: Testing the new tools. *UNSWLJ*, 37:643, 2014.

Obermeyer, Z. and Emanuel, E. J. Predicting the futurebig data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13):1216, 2016.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5680–5689. 2017.

Shaw, M. J. and Gentry, J. A. Using an expert system with inductive learning to evaluate business loans. *Financial Management*, pp. 45–56, 1988.

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. Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine*, 8(1): 68, 2002.

Singh, A. and Joachims, T. Fairness of exposure in rankings. *arXiv preprint arXiv:1802.07281*, 2018.

Sion, M. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

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

Von Neumann, J. and Morgenstern, O. Theory of games and economic behavior. *Bull. Amer. Math. Soc*, 51(7): 498–504, 1945.

Wightman, L. F. Lsac national longitudinal bar passage study. 1998.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *Proceedings of the International Conference on Machine Learning*, pp. III–325–III–333. JMLR.org, 2013.

Appendix A. Proofs

A.1 Proof of Theorem 1

Proof of Theorem 1. Using the minimax duality we perform the following transformations:

$$\min_{\widehat{P}\in\Delta\cap\Gamma}\ \max_{\check{P}\in\Delta\cap\Xi}\ \mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y)\\ \check{P}(\widehat{y}|\mathbf{x},a,y)}}\left[-\log\widehat{P}(\widehat{Y}|\mathbf{X},A,Y)\right] \tag{28}$$

$$\overset{(a)}{=}\max_{\check{P}\in\Delta\cap\Xi}\ \min_{\widehat{P}\in\Delta\cap\Gamma}\ \mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y)\\ \check{P}(\widehat{y}|\mathbf{x},a,y)}}\left[-\log\widehat{P}(\widehat{Y}|\mathbf{X},A,Y)\right] \tag{29}$$

$$\overset{(b)}{=}\max_{\check{P}\in\Delta}\min_{\theta}\ \min_{\widehat{P}\in\Delta\cap\Gamma}\ \mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y)\\ \check{P}(\widehat{y}|\mathbf{x},a,y)}}\left[-\log\widehat{P}(\widehat{Y}|\mathbf{X},A,Y)\right]+\theta^\top\Big(\mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y)\\ \check{P}(\widehat{y}|\mathbf{x},a,y)}}[\phi(\mathbf{X},\widehat{Y})]-\mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)}[\phi(\mathbf{X},Y)]\Big) \tag{30}$$

$$\overset{(c)}{=}\min_{\theta}\max_{\check{P}\in\Delta}\ \min_{\widehat{P}\in\Delta\cap\Gamma}\ \mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y)\\ \check{P}(\widehat{y}|\mathbf{x},a,y)}}\left[-\log\widehat{P}(\widehat{Y}|\mathbf{X},A,Y)\right]+\theta^\top\Big(\mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y)\\ \check{P}(\widehat{y}|\mathbf{x},a,y)}}[\phi(\mathbf{X},\widehat{Y})]-\mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)}[\phi(\mathbf{X},Y)]\Big) \tag{31}$$

$$\overset{(d)}{=}\min_{\theta}\ \min_{\widehat{P}\in\Delta\cap\Gamma}\ \max_{\check{P}\in\Delta}\mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y)\\ \check{P}(\widehat{y}|\mathbf{x},a,y)}}\left[-\log\widehat{P}(\widehat{Y}|\mathbf{X},A,Y)\right]+\theta^\top\Big(\mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y)\\ \check{P}(\widehat{y}|\mathbf{x},a,y)}}[\phi(\mathbf{X},\widehat{Y})]-\mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)}[\phi(\mathbf{X},Y)]\Big) \tag{32}$$

$$\overset{(e)}{=}\min_{\theta}\min_{\widehat{P}\in\Delta}\max_{\lambda}\max_{\check{P}\in\Delta}\mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y)\\ \check{P}(\widehat{y}|\mathbf{x},a,y)}}\left[-\log\widehat{P}(\widehat{Y}|\mathbf{X},A,Y)\right]+\theta^\top\Big(\mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y)\\ \check{P}(\widehat{y}|\mathbf{x},a,y)}}[\phi(\mathbf{X},\widehat{Y})]-\mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)}[\phi(\mathbf{X},Y)]\Big) \tag{33}$$

$$+\lambda\Big(\tfrac{1}{p_{\gamma_1}}\mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y)\\ \widehat{P}(\widehat{y}|\mathbf{x},a,y)}}[\mathbb{I}(\widehat{Y}=1\wedge\gamma_1(A,Y)]-\tfrac{1}{p_{\gamma_0}}\mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y)\\ \widehat{P}(\widehat{y}|\mathbf{x},a,y)}}[\mathbb{I}(\widehat{Y}=1\wedge\gamma_0(A,Y))]\Big)$$

$$\overset{(f)}{=}\min_{\theta}\max_{\lambda}\min_{\widehat{P}\in\Delta}\max_{\check{P}\in\Delta}\mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y)\\ \check{P}(\widehat{y}|\mathbf{x},a,y)}}\left[-\log\widehat{P}(\widehat{Y}|\mathbf{X},A,Y)\right]+\theta^\top\Big(\mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y)\\ \check{P}(\widehat{y}|\mathbf{x},a,y)}}[\phi(\mathbf{X},\widehat{Y})]-\mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)}[\phi(\mathbf{X},Y)]\Big) \tag{34}$$

$$+\lambda\Big(\tfrac{1}{p_{\gamma_1}}\mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y)\\ \widehat{P}(\widehat{y}|\mathbf{x},a,y)}}[\mathbb{I}(\widehat{Y}=1\wedge\gamma_1(A,Y))]-\tfrac{1}{p_{\gamma_0}}\mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y)\\ \widehat{P}(\widehat{y}|\mathbf{x},a,y)}}[\mathbb{I}(\widehat{Y}=1\wedge\gamma_0(A,Y))]\Big)$$

$$\overset{(g)}{=}\min_{\theta}\max_{\lambda}\min_{\widehat{P}\in\Delta}\max_{\check{P}\in\Delta}\mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)}\Bigg[\mathbb{E}_{\check{P}(\widehat{y}|\mathbf{x},a,y)}\left[-\log\widehat{P}(\widehat{Y}|\mathbf{X},A,Y)\right]+\theta^\top\Big(\mathbb{E}_{\check{P}(\widehat{y}|\mathbf{x},a,y)}[\phi(\mathbf{X},\widehat{Y})]-\phi(\mathbf{X},Y)\Big) \tag{35}$$

$$+\lambda\Big(\tfrac{1}{p_{\gamma_1}}\mathbb{E}_{\widehat{P}(\widehat{y}|\mathbf{x},a,y)}[\mathbb{I}(\widehat{Y}=1\wedge\gamma_1(A,Y))]-\tfrac{1}{p_{\gamma_0}}\mathbb{E}_{\widehat{P}(\widehat{y}|\mathbf{x},a,y)}[\mathbb{I}(\widehat{Y}=1\wedge\gamma_0(A,Y))]\Big)\Bigg]$$

The transformation steps above are described as follows:

(a) We flip the min and max order using minimax duality (Von Neumann & Morgenstern, 1945). The domains of $\widehat{P}$ and $\check{P}$ are both compact convex sets and the objective function is convex over $\widehat{P}$ and concave over $\check{P}$, therefore, strong duality holds.

(b) We introduce the Lagrange dual variable θ to directly incorporate the moment matching constraints over $\check{P}$ into the objective function.

(c) The domain of $\check{P}$ is a compact convex subset of $\mathbb{R}^n$, while the domain of θ is $\mathbb{R}^m$. The objective is concave on $\check{P}$ for all θ, while it is convex on θ for all $\check{P}$. Based on Sion's minimax theorem (Sion, 1958), strong duality holds, and thus we can flip the optimization order of $\check{P}$ and θ.

(d) We flip the inner min and max over $\widehat{P}$ and $\check{P}$ using the minimax duality, as in (a).

(e) We introduce the Lagrange dual variable λ to directly incorporate the fairness constraints over $\widehat{P}$ into the objective function.

(f) Similar to (c), we use Sion's minimax theorem to flip the optimization order of λ and $\widehat{P}$.

(g) We group the expectation with respect to the empirical training data.

We now focus on the inner minimax formulation over $\widehat{P}$ and $\widecheck{P}$, given the value of θ and λ, i.e.:

$$\min_{\widehat{P}\in\Delta}\max_{\widecheck{P}\in\Delta}\ \mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)}\Bigg[\mathbb{E}_{\widecheck{P}(\widehat{y}|\mathbf{x},a,y)}\Big[-\log\widehat{P}(\widehat{Y}|\mathbf{X},A,Y)\Big]+\theta^\top\Big(\mathbb{E}_{\widecheck{P}(\widehat{y}|\mathbf{x},a,y)}[\phi(\mathbf{X},\widehat{Y})]-\phi(\mathbf{X},Y)\Big) \\ +\lambda\Big(\tfrac{1}{p_{\gamma_1}}\mathbb{E}_{\widehat{P}(\widehat{y}|\mathbf{x},a,y)}[\mathbb{I}(\widehat{Y}=1\wedge\gamma_1(A,Y))]-\tfrac{1}{p_{\gamma_0}}\mathbb{E}_{\widehat{P}(\widehat{y}|\mathbf{x},a,y)}[\mathbb{I}(\widehat{Y}=1\wedge\gamma_0(A,Y))]\Big)\Bigg] \tag{36}$$

We aim to find the analytical solution for $\widehat{P}$ and $\widecheck{P}$ in the equation above. First, we write the Lagrangian by incorporating the probability simplex constraints into the objective, i.e.:

$$\min_{\widehat{P}}\max_{\widecheck{P}}\min_{\widecheck{\alpha},\beta\geq 0}\max_{\widehat{\alpha}}L(\widehat{P},\widecheck{P},\widehat{\alpha},\widecheck{\alpha},\beta)=\min_{\widecheck{\alpha},\beta\geq 0}\max_{\widehat{\alpha}}\max_{\widecheck{P}}\min_{\widehat{P}}\mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)\widecheck{P}(\widehat{y}|\mathbf{x},a,y)}[-\log(\widehat{P}(\widehat{Y}|\mathbf{X},A,Y))] \\ +\theta^\top(\mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)\widecheck{P}(\widehat{y}|\mathbf{x},a,y)}[\phi(\mathbf{X},\widehat{Y})]-\mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)}[\phi(\mathbf{X},Y)])+\mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)\widehat{P}(\widehat{y}|\mathbf{x},a,y)}[F_\lambda(A,Y,\widehat{Y})] \\ +\sum_{(\mathbf{x},a,y)\in\mathcal{D}}\widehat{\alpha}(\mathbf{x},a)\Big[\mathbb{E}_{\widehat{P}(\widehat{y}|\mathbf{x},a,y)}[1|\mathbf{x},a,y]-1\Big]+\sum_{(\mathbf{x},a,y)\in\mathcal{D}}\widecheck{\alpha}(\mathbf{x},a)\Big[\mathbb{E}_{\widecheck{P}(\widehat{y}|\mathbf{x},a,y)}[1|\mathbf{x},a,y]-1\Big] \\ +\sum_{\mathbf{x},a,y\in\mathcal{D}}\sum_{\widehat{y}\in\mathcal{Y}}\beta(\mathbf{x},a,y,\widehat{y})\widecheck{P}(\widehat{y}|\mathbf{x},a,y) \tag{37}$$

$$\text{where : }F_\lambda(a,y,\widehat{y})=\begin{cases}\frac{\lambda}{p_{\gamma_1}} & \text{if }\widehat{y}=1\wedge\gamma_1(a,y)\\ -\frac{\lambda}{p_{\gamma_0}} & \text{if }\widehat{y}=1\wedge\gamma_0(a,y)\\ 0 & \text{otherwise.}\end{cases} \tag{38}$$

We now take the derivative of the Lagrangian with respect to $\widehat{P}(\hat{y}|\mathbf{x},a,y)$:

$$\frac{\partial L}{\partial\widehat{P}(\hat{y}|\mathbf{x},a,y)}=-\frac{\widetilde{P}(\mathbf{x},a,y)\widecheck{P}(\hat{y}|\mathbf{x},a,y)}{\widehat{P}(\hat{y}|\mathbf{x},a,y)}+\widetilde{P}(\mathbf{x},a,y)F_\lambda(a,y,\hat{y})+\hat{\alpha}(x,a). \tag{39}$$

By setting Eq. (39) to zero, we rewrite $\widehat{P}$ in terms of $\widecheck{P}$:

$$\widehat{P}(\hat{y}|\mathbf{x},a,y)=\frac{\widetilde{P}(\mathbf{x},a,y)\widecheck{P}(\hat{y}|\mathbf{x},a,y)}{\widetilde{P}(\mathbf{x},a,y)F_\lambda(a,y,\hat{y})+\hat{\alpha}(x,a)}=\frac{\widecheck{P}(\hat{y}|\mathbf{x},a,y)}{F_\lambda(a,y,\hat{y})+\frac{\hat{\alpha}(x,a)}{\widetilde{P}(\mathbf{x},a,y)}}. \tag{40}$$

Using Eq. (39) we rewrite Eq. (37) as:

$$L(\widecheck{P},\widehat{\alpha},\widecheck{\alpha},\beta)=\min_{\widecheck{\alpha},\beta\geq 0}\max_{\widehat{\alpha}}\max_{\widecheck{P}}\mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)\widecheck{P}(\widehat{y}|\mathbf{x},a,y)}\Big[-\log(\widehat{P}(\widehat{Y}|\mathbf{X},A,Y))\Big] \\ +\theta^\top(\mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)\widecheck{P}(\widehat{y}|\mathbf{x},a,y)}\Big[\phi(\mathbf{X},\widehat{Y})\Big]-\mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)}[\phi(\mathbf{X},Y)]) \\ \underbrace{+\mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)\widehat{P}(\widehat{y}|\mathbf{x},a,y)}\Big[F_\lambda(A,Y,\widehat{Y})\Big]+\mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)\widehat{P}(\widehat{y}|\mathbf{x},a,y)}\left[\frac{\widehat{\alpha}(\mathbf{x},a)}{\widetilde{P}(\mathbf{x},a,y)}\right]}_{=\mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)\check{P}(\widehat{y}|\mathbf{x},a,y)}[1]=1}-\mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)}\left[\frac{\widehat{\alpha}(\mathbf{x},a)}{\widetilde{P}(\mathbf{x},a,y)}\right] \\ +\mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)\widecheck{P}(\widehat{y}|\mathbf{x},a,y)}\left[\frac{\widecheck{\alpha}(\mathbf{x},a)}{\widetilde{P}(\mathbf{x},a,y)}\right]-\mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)}\left[\frac{\widecheck{\alpha}(\mathbf{x},a)}{\widetilde{P}(\mathbf{x},a,y)}\right]+\mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)\widecheck{P}(\widehat{y}|\mathbf{x},a,y)}\left[\frac{\beta(\mathbf{x},a,y,\widehat{y})}{\widetilde{P}(\mathbf{x},a,y)}\right]. \tag{41}$$

Replacing $\widehat{P}$ in Lagrangian we get:

$$
\begin{aligned}
L(\check{P}, \widehat{\alpha}, \breve{\alpha}, \beta) = \min_{\breve{\alpha}, \beta \geq 0} \max_{\widehat{\alpha}} \max_{\check{P}} \mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)\check{P}(\widehat{y}|\mathbf{x},a,y)} &\left[- \log \check{P}(\widehat{Y}|\mathbf{X}, A, Y) + \log \left(F_\lambda(a, y, \hat{y}) + \frac{\hat{\alpha}(x, a)}{\widetilde{P}(\mathbf{x}, a, y)} \right) \right] \\
&+ \theta^\top (\mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)\check{P}(\widehat{y}|\mathbf{x},a,y)} \left[\phi(\mathbf{X}, \widehat{Y}) \right] - \mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)} \left[\phi(\mathbf{X}, Y) \right]) + 1 - \mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)} \left[\frac{\widehat{\alpha}(\mathbf{x}, a)}{\widetilde{P}(\mathbf{x}, a, y)} \right] \\
&+ \mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)\check{P}(\widehat{y}|\mathbf{x},a,y)} \left[\frac{\breve{\alpha}(\mathbf{x}, a)}{\widetilde{P}(\mathbf{x}, a, y)} \right] - \mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)} \left[\frac{\breve{\alpha}(\mathbf{x}, a)}{\widetilde{P}(\mathbf{x}, a, y)} \right] + \mathbb{E}_{\widetilde{P}(\mathbf{x},a,y)\check{P}(\widehat{y}|\mathbf{x},a,y)} \left[\frac{\beta(\mathbf{x}, a, y, \widehat{y})}{\widetilde{P}(\mathbf{x}, a, y)} \right].
\end{aligned}
\tag{42}
$$

We now calculate the derivative with respect to $\check{P}$.

$$
\frac{\partial L}{\partial \check{P}(\widehat{y}|\mathbf{x}, a, y)} = \widetilde{P}(\mathbf{x}, a, y) \left(- \log(\check{P}(\widehat{y}|\mathbf{x}, a, y)) + \theta \phi(\mathbf{x}, \hat{y}) + \log \left(F_\lambda(a, y, \hat{y}) + \frac{\hat{\alpha}(x_i, a)}{\widetilde{P}(\mathbf{x}, a, y)} \right) + \frac{\breve{\alpha}(\mathbf{x}, a)}{\widetilde{P}(\mathbf{x}, a, y)} + \frac{\beta(\mathbf{x}, a, y, \widehat{y})}{\widetilde{P}(\mathbf{x}, a, y)} \right)
\tag{43}
$$

Setting Eq. (43) to 0 yields:

$$
\log \left(\frac{\check{P}(\widehat{y}|\mathbf{x}, a, y)}{F_\lambda(a, y, \hat{y}) + \frac{\hat{\alpha}(x_i, a)}{\widetilde{P}(\mathbf{x}, a, y)}} \right) = \theta \phi(\mathbf{x}, \hat{y}) + \frac{\breve{\alpha}(\mathbf{x}, a)}{\widetilde{P}(\mathbf{x}, a, y)} + \frac{\beta(\mathbf{x}, a, y, \widehat{y})}{\widetilde{P}(\mathbf{x}, a, y)}
\tag{44}
$$

$$
\widehat{P}(\widehat{y}|\mathbf{x}, a, y) = e^{\theta^\top \phi(\mathbf{x}, \hat{y}) + \frac{\breve{\alpha}(\mathbf{x}, a)}{\widetilde{P}(\mathbf{x}, a, y)} + \frac{\beta(\mathbf{x}, a, y, \hat{y})}{\widetilde{P}(\mathbf{x}, a, y)}}.
\tag{45}
$$

We analytically solve the normalization constraint for $\widehat{P}$, i.e., $\sum_{\widehat{y} \in \mathcal{Y}} \widehat{P}(\hat{y}|\mathbf{x}, a, y) = 1$

$$
\sum_{\widehat{y} \in \mathcal{Y}} e^{\theta^\top \phi(\mathbf{x}, \hat{y}) + \frac{\breve{\alpha}(\mathbf{x}, a)}{\widetilde{P}(\mathbf{x}, a, y)} + \frac{\beta(\mathbf{x}, a, y, \hat{y})}{\widetilde{P}(\mathbf{x}, a, y)}} = 1
\tag{46}
$$

$$
\frac{\breve{\alpha}(\mathbf{x}, a)}{\widetilde{P}(\mathbf{x}, a, y)} = - \log \left(\sum_{\hat{y} \in \mathcal{Y}} e^{\theta^\top \phi(\mathbf{x}, \hat{y}) + \frac{\beta(\mathbf{x}, a, y, \hat{y})}{\widetilde{P}(\mathbf{x}, a, y)}} \right),
\tag{47}
$$

which yields following parametric form of the predictor distribution:

$$
\widehat{P}(\widehat{y}|\mathbf{x}, a, y) = \frac{e^{\theta^\top \phi(\mathbf{x}, \hat{y}) + \frac{\beta(\mathbf{x}, a, y, \widehat{y})}{\widetilde{P}(\mathbf{x}, a, y)}}}{Z_\theta(\mathbf{x}, a, y)} = \frac{e^{\theta^\top \phi(\mathbf{x}, \widehat{y}) + \frac{\beta(\mathbf{x}, a, y, \widehat{y})}{\widetilde{P}(\mathbf{x}, a, y)}}}{\sum_{y' \in \mathcal{Y}} e^{\theta^\top \phi(\mathbf{x}, y') + \frac{\beta(\mathbf{x}, a, y, y')}{\widetilde{P}(\mathbf{x}, a, y)}}}.
\tag{48}
$$

Notice the similarity to standard logistic regression. Where in contrast, here the probability for each class is adjusted with terms $\frac{\beta(\mathbf{x}, a, y, \hat{y})}{\widetilde{P}(\mathbf{x}, a, y)}$ to satisfy the fairness constraints.

From Eq. (40) we get the relation of $\check{P}$ and $\widehat{P}$. Solving the normalization constraint for $\check{P}(\widehat{y}|\mathbf{x}, a, y)$ yields:

$$
\sum_{\widehat{y} \in \mathcal{Y}} \check{P}(\widehat{y}|\mathbf{x}, a, y) = 1
\tag{49}
$$

$$
\sum_{\widehat{y} \in \mathcal{Y}} \widehat{P}(\widehat{y}|\mathbf{x}, a, y) \left(F_\lambda(a, y, \widehat{y}) + \frac{\widehat{\alpha}(\mathbf{x}, a)}{\widetilde{P}(\mathbf{x}, a, y)} \right) = 1
\tag{50}
$$

$$
\frac{\widehat{\alpha}(\mathbf{x}, a)}{\widetilde{P}(\mathbf{x}, a, y)} = 1 - \sum_{\widehat{y} \in \mathcal{Y}} \widehat{P}(\widehat{y}|\mathbf{x}, a, y) F_\lambda(a, y, \widehat{y})
\tag{51}
$$

Thus, we can rewrite $\check{P}$ as:

$$
\check{P}(\widehat{y}|\mathbf{x}, a, y) = \widehat{P}(\widehat{y}|\mathbf{x}, a, y)(F_\lambda(a, y, \widehat{y}) + 1 - \sum_{\bar{y} \in \mathcal{Y}} \widehat{P}(y'|\mathbf{x}, a, y) F_\lambda(a, y, y'))
\tag{52}
$$

We consider the binary classification $\widehat{y}, y = \{0,1\}$, and expand $\check{P}$ as:

$$\begin{aligned}
\check{P}(\widehat{y}=1|\mathbf{x},a,y) &= \widehat{P}(\widehat{y}=1|\mathbf{x},a,y)[F_\lambda(a,y,1)+1-\widehat{P}(\widehat{y}=1|\mathbf{x},a,y)F_\lambda(a,y,1))] && (53)\\
&= \widehat{P}(\widehat{y}=1|\mathbf{x},a,y)(1+\widehat{P}(\widehat{y}=0|\mathbf{x},a,y)F_\lambda(a,y,1)) && (54)\\
&= \begin{cases} \widehat{P}(\widehat{y}=1|\mathbf{x},a,y)(1+\frac{\lambda}{p_{\gamma_1}}\widehat{P}(\widehat{y}=0|\mathbf{x},a,y)) & \text{if } \gamma_1(a,y)\\ \widehat{P}(\widehat{y}=1|\mathbf{x},a,y)(1-\frac{\lambda}{p_{\gamma_0}}\widehat{P}(\widehat{y}=0|\mathbf{x},a,y)) & \text{if } \gamma_0(a,y)\\ \widehat{P}(\widehat{y}=1|\mathbf{x},a,y) & \text{otherwise} \end{cases} && (55)
\end{aligned}$$

The above equation shows that the adversary distribution $\check{P}$ is a quadratic function of predictor $\widehat{P}$, for example in the case where $\gamma_1(a,y)=1$:

$$\check{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y) = \rho(1+\tfrac{\lambda}{p_{\gamma_1}}(1-\rho)) = (1+\tfrac{\lambda}{p_{\gamma_1}})\rho - \tfrac{\lambda}{p_{\gamma_1}}\rho^2,$$

where $\rho = \widehat{P}_{\theta,\lambda}(\widehat{y}{=}1|\mathbf{x},a,y)$. For the region where the function goes above 1 (or below 1 depending on sign of F_λ), the predictor probability must be truncated in terms of fairness function such that $\check{P}=1$ (or zero). We derive these cases in the following by considering that the complementary slackness ensures positivity of $\check{P}$.

The complementary slackness from KKT condition, requires:

$$\forall_{\mathbf{x},a,y,\widehat{y}} \quad \beta(\mathbf{x},a,y,\widehat{y})\check{P}(\widehat{y}|\mathbf{x},a,y) = 0 \tag{56}$$

Suppose the case $\check{P}(\widehat{y}|\mathbf{x},a,y)=0$.

$$\begin{aligned}
\check{P}(\widehat{y}|\mathbf{x},a,y) &= \widehat{P}(\widehat{y}|\mathbf{x},a,y)(F_\lambda(a,y,\hat{y})+1-\sum_{\bar{y}\in\mathcal{Y}}\widehat{P}(\bar{y}|\mathbf{x},a,y)F_\lambda(a,y,\bar{y}))=0 && (57)\\
\widehat{P}>0 &\implies F_\lambda(a,y,\widehat{y})+1-\sum_{\bar{y}\in\{0,1\}}\widehat{P}(\bar{y}|\mathbf{x},a,y)F_\lambda(a,y,\bar{y})=0 && (58)\\
& F_\lambda(a,y,\widehat{y})+1-\widehat{P}(0|\mathbf{x},a,y)F_\lambda(a,y,0)-\widehat{P}(1|\mathbf{x},a,y)F_\lambda(a,y,1)=0 && (59)
\end{aligned}$$

Since $F_\lambda(a,y,0)=0$, then the equation above reduces to:

$$F_\lambda(a,y,\widehat{y})+1-\widehat{P}(\widehat{y}=1|\mathbf{x},a,y)F_\lambda(a,y,1)=0 \tag{60}$$

$$\widehat{P}(\widehat{y}=1|\mathbf{x},a,y) = \frac{F_\lambda(a,y,\widehat{y})+1}{F_\lambda(a,y,1)} \tag{61}$$

Observe that the above equation can only hold if $\gamma_1(a,y)=1$, or $\gamma_0(a,y)=1$. For the other cases, complementary slackness requires that $\beta(\mathbf{x},a,y,\widehat{y})=0$ and $\check{P}(\widehat{y}|\mathbf{x},a,y)=\widehat{P}(\widehat{y}|\mathbf{x},a,y)=\frac{e^{\theta^\top\phi(\mathbf{x},\widehat{y})}}{Z_\theta(\mathbf{x})}$

Thus, we have the following cases:

$$\widehat{P}(\widehat{y}=1|\mathbf{x},a,y) = \begin{cases} \frac{p_{\gamma_1}}{\lambda} & \text{if } \gamma_1(a,y)\wedge\check{P}(1|\mathbf{x},a,y)=1,\\ -\frac{p_{\gamma_0}}{\lambda} & \text{if } \gamma_0(a,y)\wedge\check{P}(1|\mathbf{x},a,y)=1\\ 1+\frac{p_{\gamma_1}}{\lambda} & \text{if } \gamma_1(a,y)\wedge\check{P}(1|\mathbf{x},a,y)=0\\ 1-\frac{p_{\gamma_0}}{\lambda} & \text{if } \gamma_0(a,y)\wedge\check{P}(1|\mathbf{x},a,y)=0\\ \frac{e^{\theta^\top\phi(\mathbf{x},1)}}{Z_\theta(\mathbf{x})} & \text{otherwise} \end{cases} \tag{62}$$

Therefore, if $\lambda\geq 0$, we have following parametric form for the predictor distribution:

$$\widehat{P}(\widehat{y}=1|\mathbf{x},a,y) = \begin{cases} \min\{\frac{p_{\gamma_1}}{\lambda},\frac{e^{\theta^\top\phi(\mathbf{x},1)}}{Z_\theta(\mathbf{x})}\} & \text{if } \gamma_1(a,y)\\ \max\{1-\frac{p_{\gamma_0}}{\lambda},\frac{e^{\theta^\top\phi(\mathbf{x},1)}}{Z_\theta(\mathbf{x})}\} & \text{if } \gamma_0(a,y)\\ \frac{e^{\theta\phi(\mathbf{x},1)}}{Z_\theta(\mathbf{x})} & \text{otherwise} \end{cases} \tag{63}$$

and if $\lambda < 0$:

$$\widehat{P}(\widehat{y} = 1|\mathbf{x}, a, y) = \begin{cases} \max\{1 + \frac{p_{\gamma_1}}{\lambda}, \frac{e^{\theta^\top \phi(\mathbf{x},1)}}{Z_\theta(\mathbf{x})}\} & \text{if } \gamma_1(a, y) \\ \min\{-\frac{p_{\gamma_1}}{\lambda}, \frac{e^{\theta^\top \phi(\mathbf{x},1)}}{Z_\theta(\mathbf{x})}\} & \text{if } \gamma_0(a, y) \\ \frac{e^{\theta\phi(\mathbf{x},1)}}{Z_\theta(\mathbf{x})} & \text{otherwise} \end{cases} \cdot \tag{64}$$

Note that if $\lambda = 0$, all of the cases collapse to a single case $\widehat{P}(\widehat{y} = 1|\mathbf{x}, a, y) = \frac{e^{\theta\phi(\mathbf{x},1)}}{Z_\theta(\mathbf{x})}$.

□

A.2 Proof of Theorem 2

Proof of Theorem 2. Given the optimum λ^*_θ for each θ, Eq. (11) reduces to:

$$\mathcal{L} = \min_\theta \frac{1}{n} \sum_{(\mathbf{x},a,y)\in\mathcal{D}} \left\{ \mathbb{E}_{\check{P}_{\theta,\lambda^*_\theta}(\widehat{y}|\mathbf{x},a,y)} \left[-\log \widehat{P}_{\theta,\lambda^*_\theta}(\widehat{Y}|\mathbf{x}, a, y) \right] + \theta^\top \left(\mathbb{E}_{\check{P}_{\theta,\lambda^*_\theta}(\widehat{y}|\mathbf{x},a,y)} [\phi(\mathbf{x}, \widehat{Y})] - \phi(\mathbf{x}, y) \right) \right\} \tag{65}$$

$$= \min_\theta \frac{1}{n} \sum_{(\mathbf{x},a,y)\in\mathcal{D}} \sum_{\widehat{y}\in\mathcal{Y}} -\check{P}(\widehat{y}|x, a, y) \left[\log(\widehat{P}(\widehat{y}|x, a, y)) - \theta^\top \phi(x, \widehat{y}) \right] - \theta^\top \phi(x, y) \tag{66}$$

Plugging the parametric distribution forms of $\widehat{P}$ and $\check{P}$, for $\lambda^*_\theta > 0$, we get:

$$\mathcal{L} = \min_\theta \frac{1}{n} \sum_{(\mathbf{x},a,y)\in\mathcal{D}} \begin{cases} -\log(\frac{p_{\gamma_1}}{\lambda^*_\theta}) + \theta^\top(\phi(\mathbf{x}, 1) - \phi(\mathbf{x}, y)) & \text{if } \gamma_1(a, y), \text{and } \frac{e^{\theta^\top \phi(x_i,1)}}{Z_\theta(x_i)} > \frac{p_{\gamma_1}}{\lambda^*_\theta} \\ -\log(\frac{p_{\gamma_0}}{\lambda^*_\theta}) + \theta^\top(\phi(\mathbf{x}, 0) - \phi(\mathbf{x}, y)) & \text{if } \gamma_0(a, y), \text{and } \frac{e^{\theta^\top \phi(x_i,1)}}{Z_\theta(x_i)} < 1 - \frac{p_{\gamma_0}}{\lambda^*} \\ \log(\sum_{y'\in\mathcal{Y}} e^{\theta^\top \phi(x_i,y')}) - \theta^\top \phi(x, y) & \text{otherwise,} \end{cases} \tag{67}$$

and for $\lambda^*_\theta < 0$, we get:

$$\mathcal{L} = \min_\theta \frac{1}{n} \sum_{(\mathbf{x},a,y)\in\mathcal{D}} \begin{cases} -\log(-\frac{p_{\gamma_1}}{\lambda^*_\theta}) + \theta^\top(\phi(\mathbf{x}, 0) - \phi(\mathbf{x}, y)) & \text{if } \gamma_1(a, y), \text{and } \frac{e^{\theta^\top \phi(x_i,1)}}{Z_\theta(x_i)} < 1 + \frac{p_{\gamma_1}}{\lambda^*_\theta} \\ -\log(-\frac{p_{\gamma_0}}{\lambda^*_\theta}) + \theta^\top(\phi(\mathbf{x}, 1) - \phi(\mathbf{x}, y)) & \text{if } \gamma_0(a, y), \text{and } \frac{e^{\theta^\top \phi(x_i,1)}}{Z_\theta(x_i)} > -\frac{p_{\gamma_0}}{\lambda^*} \\ \log(\sum_{y'\in\mathcal{Y}} e^{\theta^\top \phi(x_i,y')}) - \theta^\top \phi(x, y) & \text{otherwise,} \end{cases} \tag{68}$$

and for $\lambda^*_\theta = 0$, we get:

$$\mathcal{L} = \min_\theta \frac{1}{n} \sum_{(\mathbf{x},a,y)\in\mathcal{D}} \log \left(\sum_{y'\in\mathcal{Y}} e^{\theta^\top \phi(x_i,y')} \right) - \theta^\top \phi(x, y). \tag{69}$$

□

A.3 Proof of Theorem 3

Proof of Theorem 3. Taking the gradient of the objective with respect to θ, for $\lambda^*_\theta > 0$, we get:

$$\partial_\theta \mathcal{L} \ni \min_\theta \frac{1}{n} \sum_{(\mathbf{x},a,y)\in\mathcal{D}} \begin{cases} \phi(\mathbf{x}, 1) - \phi(\mathbf{x}, y) & \text{if } \gamma_1(a, y), \text{and } \frac{e^{\theta^\top \phi(x_i,1)}}{Z_\theta(x_i)} > \frac{p_{\gamma_1}}{\lambda^*_\theta} \\ \phi(\mathbf{x}, 0) - \phi(\mathbf{x}, y) & \text{if } \gamma_0(a, y), \text{and } \frac{e^{\theta^\top \phi(x_i,1)}}{Z_\theta(x_i)} < 1 - \frac{p_{\gamma_0}}{\lambda^*} \\ \sum_{y'\in\mathcal{Y}} \frac{e^{\theta^\top \phi(\mathbf{x},y')}}{Z_\theta(x)} \phi(\mathbf{x}, y') - \phi(\mathbf{x}, y) & \text{otherwise,} \end{cases} \tag{70}$$

and for $\lambda^*_\theta < 0$, we get:

$$\partial_\theta \mathcal{L} \ni \min_\theta \frac{1}{n} \sum_{(\mathbf{x},a,y)\in\mathcal{D}} \begin{cases} \phi(\mathbf{x}, 0) - \phi(\mathbf{x}, y) & \text{if } \gamma_1(a, y), \text{and } \frac{e^{\theta^\top \phi(x_i,1)}}{Z_\theta(x_i)} < 1 + \frac{p_{\gamma_1}}{\lambda^*_\theta} \\ \phi(\mathbf{x}, 1) - \phi(\mathbf{x}, y) & \text{if } \gamma_0(a, y), \text{and } \frac{e^{\theta^\top \phi(x_i,1)}}{Z_\theta(x_i)} > -\frac{p_{\gamma_0}}{\lambda^*} \\ \sum_{y'\in\mathcal{Y}} \frac{e^{\theta^\top \phi(\mathbf{x},y')}}{Z_\theta(x)} \phi(\mathbf{x}, y') - \phi(\mathbf{x}, y) & \text{otherwise,} \end{cases} \tag{71}$$

and for $\lambda_\theta^* = 0$, we get:

$$\partial_\theta \mathcal{L} \ni \min_\theta \frac{1}{n} \sum_{(\mathbf{x},a,y)\in\mathcal{D}} \sum_{y'\in\mathcal{Y}} \frac{e^{\theta^\top \phi(\mathbf{x},y')}}{Z_\theta(x)} \phi(\mathbf{x},y') - \phi(\mathbf{x},y). \tag{72}$$

This can be simplified as:

$$\partial_\theta \mathcal{L} \ni \min_\theta \frac{1}{n} \sum_{(\mathbf{x},a,y)\in\mathcal{D}} \begin{cases} \phi(\mathbf{x},1) - \phi(\mathbf{x},y) & \text{if } \gamma_1(a,y), \text{and } \frac{e^{\theta^\top \phi(x_i,1)}}{Z_\theta(x_i)} > \frac{p_{\gamma_1}}{\lambda_\theta^*}, \text{and } \lambda_\theta^* > 0, \\ & \text{or if } \gamma_0(a,y), \text{and } \frac{e^{\theta^\top \phi(x_i,1)}}{Z_\theta(x_i)} > -\frac{p_{\gamma_0}}{\lambda^*}, \text{and } \lambda_\theta^* < 0 \\ \phi(\mathbf{x},0) - \phi(\mathbf{x},y) & \text{if } \gamma_0(a,y), \text{and } \frac{e^{\theta^\top \phi(x_i,1)}}{Z_\theta(x_i)} < 1 - \frac{p_{\gamma_0}}{\lambda^*}, \text{and } \lambda_\theta^* > 0, \\ & \text{or if } \gamma_1(a,y), \text{and } \frac{e^{\theta^\top \phi(x_i,1)}}{Z_\theta(x_i)} < 1 + \frac{p_{\gamma_1}}{\lambda_\theta^*}, \text{and } \lambda_\theta^* < 0 \\ \sum_{y'\in\mathcal{Y}} \frac{e^{\theta^\top \phi(\mathbf{x},y')}}{Z_\theta(x)} \phi(\mathbf{x},y') - \phi(\mathbf{x},y) & \text{otherwise.} \end{cases} \tag{73}$$

□

A.4 Proof of Theorem 4

Proof of Theorem 4. A fully expressive feature representation constraints the adversary's distribution in our primal formulation Eq. (10) to match the true data generating distribution. Then, the optimization simplified to:

$$\begin{aligned} \widehat{P}^*(\widehat{y}|\mathbf{x},a,y) &= \operatorname*{argmin}_{\widehat{P}\in\Delta\cap\Gamma} \mathbb{E}_{P(\mathbf{x},a,y)} \left[-\log \widehat{P}(\widehat{Y}|\mathbf{X},A,Y) \right] && (74) \\ &= \operatorname*{argmin}_{\widehat{P}\in\Delta\cap\Gamma} - \sum_{(\mathbf{x},a,y)} P(\mathbf{x},a,y) \log \left(\widehat{P}(\widehat{Y} = y|\mathbf{x},a,y) \right) && (75) \\ &= \operatorname*{argmin}_{\widehat{P}\in\Delta\cap\Gamma} - \sum_{(\mathbf{x},a,y)} P(\mathbf{x},a,y) \log \left(\frac{\widehat{P}(\widehat{Y} = y|\mathbf{x},a,y)}{P(Y = y|\mathbf{x},a)} \right) - \sum_{(\mathbf{x},a,y)} P(\mathbf{x},a,y) \log \left(P(Y = y|\mathbf{x},a) \right) && (76) \\ &= \operatorname*{argmin}_{\widehat{P}\in\Delta\cap\Gamma} - \sum_{(\mathbf{x},a,y)} P(\mathbf{x},a,y) \log \left(\frac{\widehat{P}(\widehat{Y} = y|\mathbf{x},a,y)}{P(Y = y|\mathbf{x},a)} \right) && (77) \\ &= \operatorname*{argmin}_{\widehat{P}\in\Delta\cap\Gamma} \; D_{\text{KL}}(P \parallel \widehat{P}). && (78) \end{aligned}$$

This means that the optimal solution of our method when learns from the true data generating distribution with a fully expressive feature representation is the fair predicting distribution that has the minimum KL-divergence from the true distribution. □

A.5 Proof of Theorem 5

Proof of Theorem 5. For the fairness constraints that depend on the true label (e.g., E.OPP. and E.ODD.), as described in §3.5, we compute $\widehat{P}^*(\widehat{y}|\mathbf{x},a)$ using Eq. (23) with the input of $\widehat{P}^*(\widehat{y}|\mathbf{x},a,y)$ and the adversary's distribution to approximate the true distribution. Based on the proof of Theorem 4, we know that, in the limit, the adversary's distribution match with the true distribution $P(\mathbf{x},a,y)$. Hence, our prediction becomes the standard marginal probability rule (it is no longer an approximation), i.e.:

$$\widehat{P}^*(\widehat{y}|\mathbf{x},a) = \widehat{P}^*(\widehat{y}|\mathbf{x},a,y=1)P(y=1|\mathbf{x},a) + \widehat{P}^*(\widehat{y}|\mathbf{x},a,y=0)P(y=0|\mathbf{x},a). \tag{79}$$

Therefore, our predictor is the marginal predicting distribution computed from the fair predictor distribution with the closest KL-divergence from the true distribution, marginalized over the true label. □

Appendix B. Additional Experimental Results

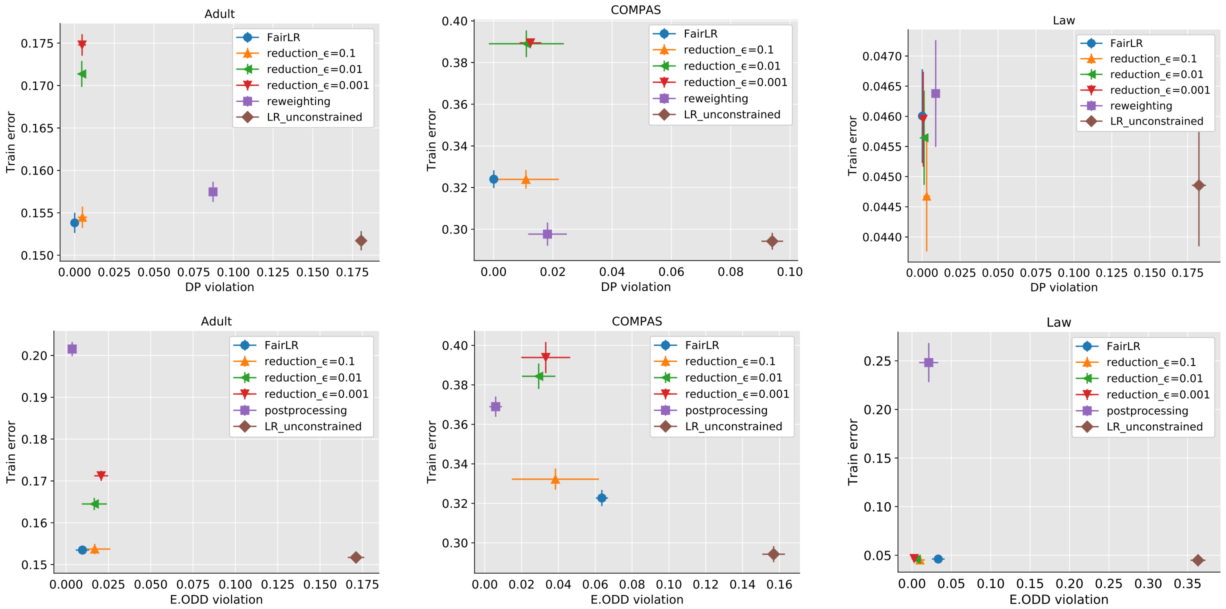

Figure 3: *Train classification error* versus *Demographic Parity* constraint violation (top rows) and *Equalized Odd* constraint violation (bottom rows). The bars indicate standard deviation on 20 random splits of data.