Fairness for Robust Log Loss Classification

Ashkan Rezaei[1*], **Rizal Fathony**[2*], **Omid Memarrast**[1], **Brian Ziebart**[1]
[1] Department of Computer Science, University of Illinois at Chicago
[2] School of Computer Science, Carnegie Mellon University
arezae4@uic.edu, rfathony@cs.cmu.edu, omemar2@uic.edu, bziebart@uic.edu

Abstract

Developing classification methods with high accuracy that also avoid unfair treatment of different groups has become increasingly important for data-driven decision making in social applications. Many existing methods enforce fairness constraints on a selected classifier (e.g., logistic regression) by directly forming constrained optimizations. We instead re-derive a new classifier from the first principles of distributional robustness that incorporates fairness criteria into a worst-case logarithmic loss minimization. This construction takes the form of a minimax game and produces a parametric exponential family conditional distribution that resembles truncated logistic regression. We present the theoretical benefits of our approach in terms of its convexity and asymptotic convergence. We then demonstrate the practical advantages of our approach on three benchmark fairness datasets.

Introduction

Though maximizing accuracy has been the principal objective for classification tasks, competing priorities are also often of key concern in practice. Fairness properties that guarantee equivalent treatment to different groups in various ways are a prime example. These may be desirable—or even legally required—when making admissions decisions for universities (Chang 2006; Kabakchieva 2013), employment and promotion decisions for organizations (Lohr 2013), medical decisions for hospitals and insurers (Shipp et al. 2002; Obermeyer and Emanuel 2016), sentencing guidelines within the judicial system (Moses and Chan 2014; O'Neil 2016), loan decisions for the financial industry (Shaw and Gentry 1988; Carter and Catlett 1987) and in many other applications. *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. We focus on three prevalent group fairness measures in this paper: *demographic parity* (Calders, Kamiran, and Pechenizkiy 2009), *equalized odds*, and *equalized opportunity* (Hardt, Price, and Srebro 2016).

[*]These two authors contributed equally.

Techniques for constructing predictors with group fairness properties can be categorized into *pre-*, *post-*, and *in-processing* methods. *Pre-processing methods* use reweighting and relabeling (Kamiran and Calders 2012; Krasanakis et al. 2018) or other transformations of input data (Calmon et al. 2017; Zemel et al. 2013; Feldman et al. 2015; Del Barrio et al. 2018; Donini et al. 2018; Zhang, Wu, and Wu 2018) to remove unfair dependencies with protected attributes. *Post-processing methods* adjust the class labels (or label distributions) provided from black box classifiers to satisfy desired fairness criteria (Hardt, Price, and Srebro 2016; Pleiss et al. 2017; Hacker and Wiedemann 2017). *In-processing methods* integrate fairness criteria into the optimization procedure of the classifier with constraints/penalties (Donini et al. 2018; Zafar et al. 2017c; 2017a; 2017b; 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), meta-algorithms (Celis et al. 2019; Menon and Williamson 2018), reduction-based methods (Agarwal 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).

Unlike many existing methods that directly form a constrained optimization from base classifiers, we take a step back and re-derive prediction from the underlying formulation of logistic regression. Working from the first principles of distributionally robust estimation (Topsøe 1979; Grünwald and Dawid 2004; Delage and Ye 2010), we incorporate fairness constraints into the formulation of the predictor. We pose predictor selection as a minimax game between a predictor that is fair on a training sample and a worst-case approximator of the training data labels that maintains some statistical properties of the training sample. Like *post-processing methods*, our approach reshapes its predictions for each group to satisfy fairness requirements. However, our approach is inherently an *in-process method* that jointly optimizes this fairness transformation and linear feature-based parameters for an exponential family distribution that can be viewed as truncated logistic regression. Our method assumes group membership attributes are given at training and testing time, which matches many real-world applications. We leave

the extension of our approach to settings with inferred group attributes as future work.

Our method reduces to a convex optimization problem with a unique solution for resolving unfairness between groups that asymptotically minimizes the KL divergence from the true distribution. In contrast, many existing methods are susceptible to the local optima of non-convex optimization or to the approximation error from relaxations (Zafar et al. 2017a; 2017c; Cotter et al. 2018; Woodworth et al. 2017; Kamishima, Akaho, and Sakuma 2011; Bechavod and Ligett 2017; Quadrianto and Sharmanska 2017), do not have unique solutions for ensuring fairness (Hardt, Price, and Srebro 2016), or produce mixtures of predictors (Agarwal et al. 2018) rather than a single coherent predictor. For fairness criteria that include the true label (e.g., equalized opportunity, equalized odds), we introduce a method for making predictions from label-conditioned distributions and establish desirable asymptotic properties. We demonstrate the practical advantages of our approach compared to existing fair classification methods on benchmark data-driven decision tasks.

Background

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: *Demographic Parity* (Calders, Kamiran, and Pechenizkiy 2009), *Equality of Opportunity* (Hardt, Price, and Srebro 2016) and *Equality of Odds* (Hardt, Price, and Srebro 2016). These are defined for binary decision settings with examples drawn from a population distribution, $(\mathbf{X}, A, Y) \sim P$, with $\widetilde{P}(\mathbf{x}, a, y)$ denoting this empirical sample distribution, $\{\mathbf{x}_i, a_i, y_i\}_{i=1:n}$. Here, $y = 1$ is the "advantaged" class for positive decisions. Each example also possesses a protected attribute $a \in \{0, 1\}$ that defines membership in one of two groups. The general decision task is to construct a probabilistic prediction, $\mathbb{P}(\widehat{y}|\mathbf{x}, a)$ over the decision variable $\widehat{y} \in \{0, 1\}$ given input $\mathbf{x} \in \boldsymbol{\mathcal{X}}$ and training data $\widetilde{P}(\mathbf{x}, a, y)$. We similarly notate an adversarial conditional distribution that approximates the labels as $\mathbb{Q}$. $\mathbb{P}$ and $\mathbb{Q}$ are the key objects being optimized in our formulation.

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 $\mathbb{P}(\widehat{y}|\mathbf{x})$, does not guarantee fairness because the protected attribute a may still be inferred from $\mathbf{x}$ (Dwork et al. 2012). Instead of imposing constraints on the predictor's inputs, definitions of fairness require statistical properties on its decisions to hold.

Definition 1. *A classifier satisfies* DEMOGRAPHIC PARITY (D.P.) *if the output variable $\widehat{Y}$ is statistically independent of the protected attribute A: $P(\widehat{Y} = 1|A = a) = P(\widehat{Y} = 1),\ \forall a \in \{0, 1\}$.*

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: $P(\widehat{Y}=1|A=a, Y=y) = P(\widehat{Y}=1|Y=y),\ \forall y, a \in \{0, 1\}$.*

Definition 3. *A classifier satisfies* EQUALIZED OPPORTUNITY (E.OPP.) *if the output variable $\widehat{Y}$ and protected attribute A are conditionally independent given $Y = 1$: $P(\widehat{Y}=1|A=a, Y=1) = P(\widehat{Y}=1|Y=1),\ \forall a \in \{0, 1\}$.*

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

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

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 specify γ_1 and γ_0 in (1) for fairness constraints (Definitions 1, 2, and 3) as:

$$\Gamma_{\text{dp}} \iff \gamma_j(A, Y) = \mathbb{I}(A = j); \quad (2)$$

$$\Gamma_{\text{e.opp}} \iff \gamma_j(A, Y) = \mathbb{I}(A = j \wedge Y = 1); \quad (3)$$

$$\Gamma_{\text{e.odd}} \iff \gamma_j(A, Y) = \begin{bmatrix} \mathbb{I}(A = j \wedge Y = 1) \\ \mathbb{I}(A = j \wedge Y = 0) \end{bmatrix}. \quad (4)$$

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

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

$$\min_{\mathbb{P}(\widehat{y}|\mathbf{x}) \in \Delta} \max_{\mathbb{Q}(\widehat{y}|\mathbf{x}) \in \Delta \cap \Xi} -\sum_{\mathbf{x},\widehat{y}} \widetilde{P}(\mathbf{x}) \mathbb{Q}(\widehat{y}|\mathbf{x}) \log \mathbb{P}(\widehat{y}|\mathbf{x}) \quad (5)$$

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

after simplifications based on the fact that the saddle point solution is $\mathbb{P} = \mathbb{Q}$. When the loss maximizer $\mathbb{Q}$ is constrained to match the statistics of training data (specified using vector-valued feature function ϕ),

$$\Xi : \Big\{\mathbb{Q} \mid \mathbb{E}_{\widetilde{P}(\mathbf{x});\mathbb{Q}(\widehat{y}|\mathbf{x})}[\phi(\mathbf{X}, \widehat{Y})] = \mathbb{E}_{\widetilde{P}(\mathbf{x},y)}[\phi(\mathbf{X}, Y)]\Big\}, \quad (6)$$

the robust log loss minimizer/maximum entropy predictor (Eq. (5)) 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 and Klein 2003). While this distribution technically needs to only be defined at input values in which training data exists (i.e., $\widetilde{P}(\mathbf{x}) > 0$), we employ an inductive assumption that generalizes the form of the distribution to other inputs.

This formulation has been leveraged to provide robust predictions under covariate shift (i.e., difference in training and testing distributions) (Liu and Ziebart 2014) and for constructing consistent predictors for multiclass classifications (Fathony et al. 2018a) and graphical models (Fathony et al. 2018b). Our approach similarly extends this fundamental formulation by imposing fairness constraints on $\mathbb{P}$. However, since the fairness constraints and statistic-matching constraints are often not fully compatible (i.e., $\Gamma \not\subseteq \Xi$), the saddle point solution is no longer simple (i.e., $\mathbb{P} \neq \mathbb{Q}$).

Formulation and Algorithms

Given fairness requirements for a predictor (Eq. (1)) and partial knowledge of the population distribution provided by a training sample (Eq. (6)), 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 a robust estimation perspective by seeking the best solution for the worst-case population distribution under these constraints.

Robust and fair log loss minimization

We formulate the robust fair predictor's construction as a minimax game between the predictor and a worst-case 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**, $\mathbb{P}$, *minimizes the worst-case log loss—as chosen by approximator* $\mathbb{Q}$ *constrained to reflect training statistics (denoted by set* Ξ *of Eq.* (6)*)—while providing empirical fairness guarantees*[1] *(denoted by set* Γ *of Eq.* (1)*):*

$$\min_{\mathbb{P}\in\Delta\cap\Gamma}\max_{\mathbb{Q}\in\Delta\cap\Xi}\mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y)\\ \mathbb{Q}(\widehat{y}|\mathbf{x},a,y)}}\Big[-\log\mathbb{P}(\widehat{Y}|\mathbf{X},A,Y)\Big]. \tag{7}$$

Though conditioning the decision variable $\widehat{Y}$ on the true label Y would appear to introduce a trivial solution ($\widehat{Y}=Y$), instead, Y only influences $\widehat{Y}$ based on fairness properties due to the robust predictor's construction. Note that if the fairness constraints do not relate Y and $\widehat{Y}$, the resulting distribution is conditionally independent (i.e., $\mathbb{P}(\widehat{Y}|\mathbf{X},A,Y=0) = \mathbb{P}(\widehat{Y}|\mathbf{X},A,Y=1)$), and when all fairness constraints are removed, this formulation reduces to the familiar logistic regression model (Manning and Klein 2003). Conveniently, this saddle point problem is convex-concave in $\mathbb{P}$ and $\mathbb{Q}$ with additional convex constraints (Γ and Ξ) on each distribution.

Parametric Distribution Form

By leveraging strong minimax duality in the "log-loss game" (Topsøe 1979; Grünwald and Dawid 2004) and strong Lagrangian duality (Boyd and Vandenberghe 2004), we derive the parametric form of our predictor.[2]

[1] Δ is the set of conditional probability simplexes (i.e., $\mathbb{P}(y|\mathbf{x},a)\geq 0, \sum_{y'}\mathbb{P}(y'|\mathbf{x},a)=1, \forall \mathbf{x},y,a$).

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

Theorem 1. *The* **Fair Robust Log-Loss Predictor** *(Definition 4) has equivalent dual formulation:*

$$\begin{aligned}\min_{\theta}\max_{\lambda}\ &\tfrac{1}{n}\sum_{(\mathbf{x},a,y)\in\mathcal{D}}\Bigg\{\mathbb{E}_{\mathbb{Q}_{\theta,\lambda}(\widehat{y}|\mathbf{x},a,y)}\Big[-\log\mathbb{P}_{\theta,\lambda}(\widehat{Y}|\mathbf{x},a,y)\Big]\\ &+\theta^\top\Big(\mathbb{E}_{\mathbb{Q}_{\theta,\lambda}(\widehat{y}|\mathbf{x},a,y)}[\phi(\mathbf{x},\widehat{Y})]-\phi(\mathbf{x},y)\Big)\\ &+\lambda\big(\tfrac{1}{p_{\gamma_1}}\mathbb{E}_{\mathbb{P}_{\theta,\lambda}(\widehat{y}|\mathbf{x},a,y)}[\mathbb{I}(\widehat{Y}=1\wedge\gamma_1(A,Y))]\\ &-\tfrac{1}{p_{\gamma_0}}\mathbb{E}_{\mathbb{P}_{\theta,\lambda}(\widehat{y}|\mathbf{x},a,y)}[\mathbb{I}(\widehat{Y}=1\wedge\gamma_0(A,Y))]\big)\Bigg\},\end{aligned} \tag{8}$$

with Lagrange multipliers θ *and* λ *for moment matching and fairness constraints, respectively, and* n *samples in the dataset. The parametric distribution of* $\mathbb{P}$ *is:*

$$\mathbb{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y)= \tag{9}$$
$$\begin{cases}\min\left\{e^{\theta^\top\phi(\mathbf{x},1)}/Z_\theta(\mathbf{x}),\frac{p_{\gamma_1}}{\lambda}\right\} & \text{if } \gamma_1(a,y)\wedge\lambda>0\\ \max\left\{e^{\theta^\top\phi(\mathbf{x},1)}/Z_\theta(\mathbf{x}),1-\frac{p_{\gamma_0}}{\lambda}\right\} & \text{if } \gamma_0(a,y)\wedge\lambda>0\\ \max\left\{e^{\theta^\top\phi(\mathbf{x},1)}/Z_\theta(\mathbf{x}),1+\frac{p_{\gamma_1}}{\lambda}\right\} & \text{if } \gamma_1(a,y)\wedge\lambda<0\\ \min\left\{e^{\theta^\top\phi(\mathbf{x},1)}/Z_\theta(\mathbf{x}),-\frac{p_{\gamma_0}}{\lambda}\right\} & \text{if } \gamma_0(a,y)\wedge\lambda<0\\ e^{\theta^\top\phi(\mathbf{x},1)}/Z_\theta(\mathbf{x}) & \text{otherwise},\end{cases}$$

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

$$\mathbb{Q}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y)=\mathbb{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y)\times \tag{10}$$
$$\begin{cases}\left(1+\frac{\lambda}{p_{\gamma_1}}\mathbb{P}_{\theta,\lambda}(\widehat{y}=0|\mathbf{x},a,y)\right) & \text{if } \gamma_1(a,y)\\ \left(1-\frac{\lambda}{p_{\gamma_0}}\mathbb{P}_{\theta,\lambda}(\widehat{y}=0|\mathbf{x},a,y)\right) & \text{if } \gamma_0(a,y)\\ 1 & \text{otherwise}.\end{cases}$$

Note that the predictor's distribution is a member of the exponential family that is similar to standard binary logistic regression, but with the option to *truncate* the probability based on the value of λ. The truncation of $\mathbb{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y)$ is from above when $0<p_{\gamma_1}/\lambda<1$ and $\gamma_1(a,y)=1$, and from below when $-1<p_{\gamma_1}/\lambda<0$ and $\gamma_1(a,y)=1$. The approximator's distribution is computed from the predictor's distribution using the quadratic function in Eq. (10), e.g., in the case where $\gamma_1(a,y)=1$:

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

where $\rho\triangleq\mathbb{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y)$. Figure 1 illustrates the relationship between $\mathbb{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y)$ and $\mathbb{Q}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y)$ for decisions influencing the fairness of group one (i.e., $\gamma_1(a,y)=1$). When $\lambda/p_{\gamma_1}=0$, the approximator's probability is equal to the predictor's probability as shown in the plot as a straight line. Positive values of λ curve the function upward (e.g., $\lambda/p_{\gamma_1}=1$) as shown in the plot. For larger λ (e.g., $\lambda/p_{\gamma_1}=2$), some of the valid predictor probabilities ($0<\mathbb{P}<1$) map to invalid approximator probabilities (i.e., $\mathbb{Q}\geq 1$) according to the quadratic function. In this case (e.g., $\lambda/p_{\gamma_1}=2$ and $\mathbb{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y)>0.5$), the predictor's probability is truncated to $p_{\gamma_1}/\lambda=0.5$ according to Eq. (9). Similarly, for negative λ, the curve is shifted downward and the predictor's probability is truncated when the quadratic function mapping results in a negative value of $\mathbb{Q}$. When

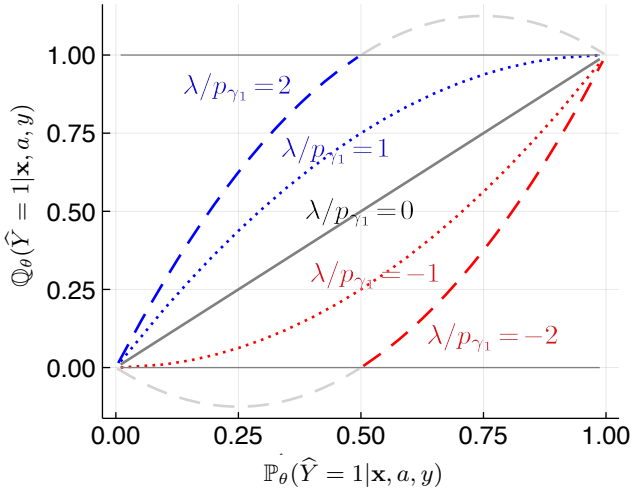

Figure 1: The relationship between predictor and approximator's distributions, $\mathbb{P}$ and $\mathbb{Q}$.

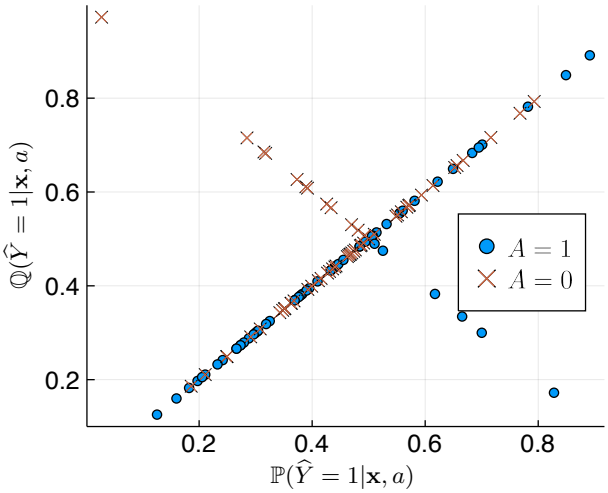

Figure 2: Post-processing correction[3] of logistic regression (Pleiss et al. 2017; Hardt, Price, and Srebro 2016) on the COMPAS dataset.

$\gamma_0(a,y)=1$, the reverse shifting is observed, i.e., shifting downward when $\lambda>0$ and shifting upward when $\lambda<0$.

We contrast our reshaping function of the decision distribution (Figure 1) with the *post-processing method* of Hardt, Price, and Srebro (2016) shown in Figure 2. Here, we use $\mathbb{Q}(\widehat{Y}=1|\mathbf{x},a)$ to represent the estimating distributions (the approximator's distribution in our method, and the standard logistic regression in Hardt, Price, and Srebro (2016)) and the post-processed predictions as $\mathbb{P}(\widehat{Y}=1|\mathbf{x},a)$. Both shift the positive prediction rates of each group to provide fairness. However, our approach provides a monotonic and parametric transformation, avoiding the criticisms that Hardt, Price, and Srebro (2016)'s modification (flipping some decisions) is partially random, creating an unrestricted hypothesis class (Bechavod and Ligett 2017). Additionally, since our parametric reshaping function is learned within an *in-processing* method, it avoids the noted suboptimalities that have been established for certain population distributions when employing post-processing alone (Woodworth et al. 2017).

Enforcing fairness constraints

The inner maximization in Eq. (8) finds the optimal λ that enforces the fairness constraint. From the perspective of the parametric distribution of $\mathbb{P}$, this is equivalent to finding threshold points (e.g., p_{γ_1}/λ and $1-p_{\gamma_0}/\lambda$) in the $\min$ and $\max$ function of Eq. (9) such that the expectation of the truncated exponential probabilities of $\mathbb{P}$ 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 points. 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 contain P_e for group γ_1 and γ_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 simultaneously. We refer to the supplementary material for the detailed algorithm.

[3]https://github.com/gpleiss/equalized_odds_and_calibration

Learning

Our learning process seeks parameters θ,λ for our distributions ($\mathbb{P}_{\theta,\lambda}$ and $\mathbb{Q}_{\theta,\lambda}$) that match the statistics of the approximator's distribution with training data (θ) and provide fairness (λ), as illustrated in Eq. (8). Using our algorithm from the previous subsection to directly compute the best λ given arbitrary values of θ, denoted λ^*_θ, the optimization of Eq. (8) reduces to 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.* (8) *reduces to:*

$$\min_\theta \tfrac{1}{n}\textstyle\sum_{(\mathbf{x},a,y)\in\mathcal{D}}\ell_{\theta,\lambda^*_\theta}(\mathbf{x},a,y), \quad \textit{where:} \tag{11}$$

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

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

Here, $T(\mathbf{x},\theta)\triangleq 1$ if the exponential probability is truncated (for example when $e^{\theta^\top\phi(\mathbf{x},1)}/Z_\theta(\mathbf{x})>p_{\gamma_1}/\lambda^*_\theta$, $\gamma_1(a,y)=1$, and $\lambda^*_\theta>0$), and is 0 otherwise.

We present an important optimization property for our objective function in the following theorem.

Theorem 3. *The objective function in Theorem 2 (Eq.* (11)*) is convex with respect to θ.*

To improve the generalizability of our parametric model, we employ a standard L2 regularization technique that is common for logistic regression models: $\theta^*=\operatorname{argmin}_\theta\sum_{(\mathbf{x},a,y)\in\mathcal{D}}\ell_{\theta,\lambda^*_\theta}(\mathbf{x},a,y)+\frac{C}{2}\|\theta\|_2^2$, where C is the

regularization constant. We employ a standard batch gradient descent optimization algorithm (e.g., L-BFGS) to obtain a solution for θ^*.[4] We also compute the corresponding solution for the inner optimization, $\lambda^*_{\theta^*}$, and then construct the optimal predictor and approximator's parametric distributions based on the values of θ^* and $\lambda^*_{\theta^*}$.

Inference

In the inference step, we apply the optimal parametric predictor distribution $\mathbb{P}_{\theta^*,\lambda^*_{\theta^*}}$ to new example inputs $(\mathbf{x}, a)$ 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. (9). Note that the predictor's parametric distribution also depends on the group membership of the example. For fairness constraints not based on the actual label Y, e.g., D.P., this parametric distribution can be directly applied to make predictions. However, for fairness constraints that depend on the true label, e.g., E.OPP. and E.ODD., we introduce a prediction procedure that estimates the true label using the approximator's parametric distribution.

For fairness constraints that depend on the true label, our algorithm outputs the predictor and approximator's parametric distributions conditioned on the value of true label, i.e., $\mathbb{P}(\widehat{y}|\mathbf{x}, a, y)$ and $\mathbb{Q}(\widehat{y}|\mathbf{x}, a, y)$. Our goal is to produce the conditional probability of $\widehat{y}$ that does not depend on the true label, i.e., $\mathbb{P}(\widehat{y}|\mathbf{x}, a)$. We construct the following procedure to estimate this probability. Based on the marginal probability rule, $\mathbb{P}(\widehat{y}|\mathbf{x}, a)$ can be expressed as:

$$\begin{aligned}\mathbb{P}(\widehat{y}|\mathbf{x}, a) = \mathbb{P}(\widehat{y}|\mathbf{x}, a, y=1)P(y=1|\mathbf{x}, a) \\ + \mathbb{P}(\widehat{y}|\mathbf{x}, a, y=0)P(y=0|\mathbf{x}, a).\end{aligned} \tag{12}$$

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

$$\begin{aligned}\mathbb{Q}(\widehat{y}|\mathbf{x}, a) \approx \mathbb{Q}(\widehat{y}|\mathbf{x}, a, y=1)\mathbb{Q}(\widehat{y}=1|\mathbf{x}, a) \\ + \mathbb{Q}(\widehat{y}|\mathbf{x}, a, y=0)\mathbb{Q}(\widehat{y}=0|\mathbf{x}, a).\end{aligned} \tag{13}$$

By rearranging the terms above, we calculate the estimate as:

$$\begin{aligned}\mathbb{Q}(\widehat{y}=1|\mathbf{x}, a) = \mathbb{Q}(\widehat{y}=1|\mathbf{x}, a, y=0)/(\mathbb{Q}(\widehat{y}=0|\mathbf{x}, a, y=1) \\ + \mathbb{Q}(\widehat{y}=1|\mathbf{x}, a, y=0)),\end{aligned} \tag{14}$$

which is directly computed from the approximator's parametric distribution produced by our model using Eq. (10). Finally, to obtain the predictor's conditional probability estimate ($\mathbb{P}(\widehat{y}|\mathbf{x}, a)$), we replace $P(y|\mathbf{x}, a)$ in Eq. (12) with $\mathbb{Q}(\widehat{y}|\mathbf{x}, a)$ calculated from Eq. (14).

Asymptotic convergence property

The ideal behavior of an algorithm is an important consideration in its design. Asymptotic convergence properties consider a learning algorithm when it is provided with access to the population 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 the population distribution $P(\mathbf{x}, a, y)$ and a fully expressive feature representation, our formulation (Def. 4) finds the* **fair** *predictor with the minimal KL-divergence from $P(\mathbf{x}, a, y)$.*

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 outputs a predictor distribution with the same desired characteristic, after being marginalized over the true label.

Theorem 5. *For fairness constraints that depend on the true label, our inference procedure in Eq.* (12) *produces the marginal predicting distribution $\mathbb{P}$ of the fair predictor distribution with the closest KL-divergence to $P(\mathbf{x}, a, y)$ in the limit.*

Experiments

Illustrative behavior on synthetic data

We illustrate the key differences between our model and logistic regression with demographic parity requirements on 2D synthetic data in Figure 3. The predictive distribution includes different truncated probabilities for each group: raising the minimum probability for group $A = 1$ and lowering the maximum probability for group $A = 0$. This permits a decision boundary that differs significantly from the logistic regression decision boundary and better realizes the desired fairness guarantees. In contrast, *post-processing methods* using logistic regression as the base classifier (Hardt, Price, and Srebro 2016) are constrained to reshape the given unfair logistic regression predictions without shifting the decision boundary orientation, often leading to suboptimality (Woodworth et al. 2017).

Datasets

We evaluate our proposed algorithm on three benchmark fairness datasets:

(1) The **UCI Adult** (Dheeru and 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, leaving 11 other features for each example.

(2) 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) and an additional nine features.

(3) The dataset from the Law School Admissions Council's National Longitudinal Bar Passage Study (Wightman 1998) has 20,649 examples. Here, the favorable outcome for the individual is passing the bar exam, with race (restricted to white and black only) as the protected attribute, and 13 other features.

[4] We refer the reader to the supplementary material for details.

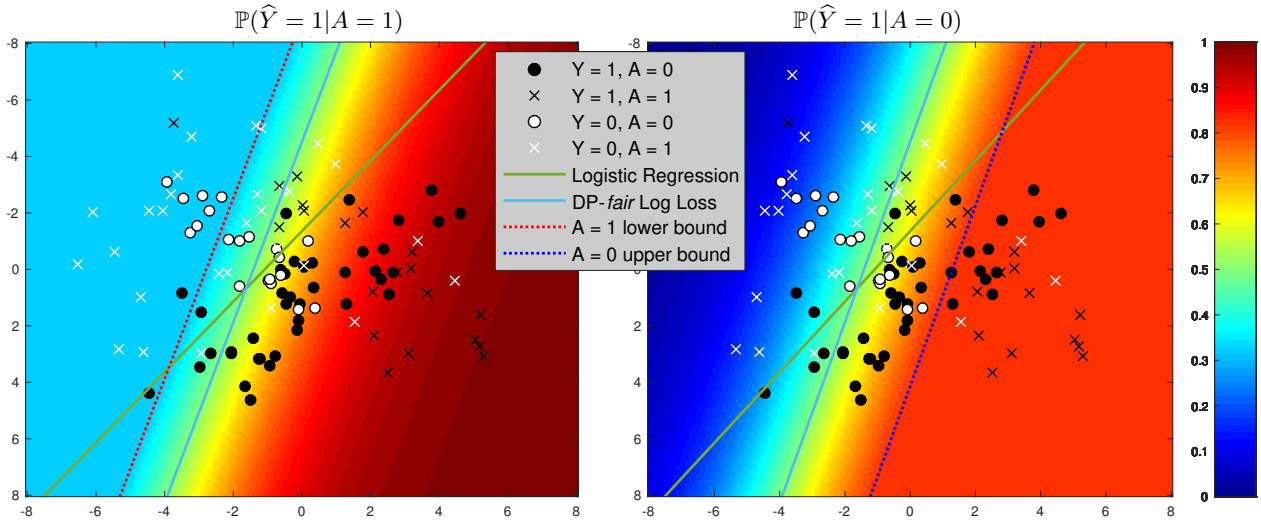

Figure 3: Experimental results on a synthetic dataset with: a heatmap indicating the predictive probabilities of our approach, along with decision and threshold boundaries; and the unfair logistic regression decision boundary.

Comparison methods

We compare our method (*Fair Log-loss*) against various baseline/fair learning algorithms that are primarily based on logistic regression as the base classifier:

(1) **Unconstrained logistic regression** is a standard logistic regression model that ignores all fairness requirements.

(2) The **cost sensitive reduction approach** 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 as the base classifier, to compare the effect of the reduction approach. We evaluate the performance of the model by varying the constraint bounds across the set $\epsilon \in \{.001, .01, .1\}$.

(3) The **constraint-based learning method**[5] of (Zafar et al. 2017c; 2017a) uses a covariance proxy measure to achieve equalized odds (under the name disparate mistreatment) (Zafar et al. 2017a), and improve the disparate impact ratio (Zafar et al. 2017c), which we use as a baseline method to evaluate demographic parity violation. They cast the resulting non-convex optimization as a disciplined convex-concave program in training time. We use the logistic regression as the base classifier.

(4) For demographic parity, we compare with the **reweighting method** (*reweighting*) of Kamiran and 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 IBM AIF360 toolkit to run this method.

(5) For equalized odds, we also compare with the **post-processing method** of Hardt, Price, and Srebro (2016) which transforms the classifier's output by solving a linear program that finds a prediction minimizing misclassification errors and satisfying 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).

[5]https://github.com/mbilalzafar/fair-classification

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, $\mathbb{E}[\widehat{Y} \neq Y]$) 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: $\big|\mathbb{E}[\mathbb{I}(\widehat{Y}=1)|A=1] - \mathbb{E}[\mathbb{I}(\widehat{Y}=1)|A=0]\big|$, and the sum of fairness violations for each class to measure the total violation for equalized odds (E.ODD.) as: $\sum_{y\in\{0,1\}} \big(\big|\mathbb{E}[\mathbb{I}(\widehat{Y}=1)|A=1,Y=y] - \mathbb{E}[\mathbb{I}(\widehat{Y}=1)|A=0,Y=y]\big|\big)$, to obtain a level comparison across different methods. We follow the methodology of Agarwal et al. (2018) to give all methods access to the protected attribute both at training and testing time by including the protected attribute in the feature vector. 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 ($\{.001, .005, .01, .05, .1, .2, .3, .4, .5\}$).

Experimental Results

Figure 4 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 Fairness can be vacuously achieved by an agnostic predictor that always outputs labels according to independent (biased)

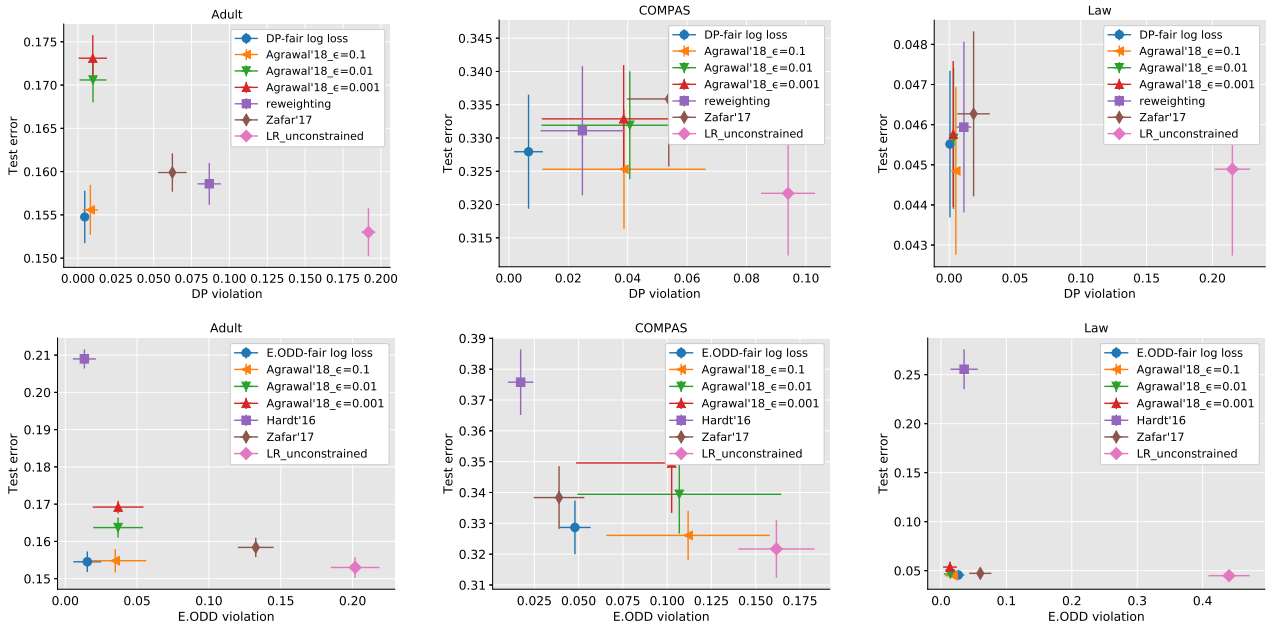

Figure 4: *Test classification error* versus *Demographic Parity* (top row) and *Equalized Odds* (bottom row) constraint violations. The bars indicate standard deviation on 20 random splits of data.

coin flips. 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 logistic regression model for how much of an increase in fairness?"

For demographic parity on the Adult dataset, our *Fair Log-loss* approach outperforms all baseline methods on average for both test error rate and for fairness violation, and on COMPAS dataset it achieves the lowest ratio of increased fairness over increased error. 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 *Fair Log-loss* approach still resides in the Pareto optimal set, i.e., there are no other methods that are significantly better than our result on both criteria. For equalized odds, *Fair Log-loss* provides the lowest ratios of increased fairness over increased 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. This shows that the approximation in our prediction procedure does not significantly impact the performance of our method. In terms of the running time, our method is an order of magnitude faster than comparable methods (e.g., the train and test running time on one random split of the Adult dataset takes approximately 5 seconds by our algorithm, 80 seconds for the constraint-based method (Zafar et al. 2017c), and 100 seconds for the reduction-based method (Agarwal et al. 2018)).

Conclusions and Future Work

We have developed a novel approach for providing fair data-driven decision making in this work by deriving a new classifier from the first principles of distributionally robust estimation (Topsøe 1979; Grünwald and Dawid 2004; Delage and Ye 2010). We formulated a learning objective 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. This resulted in a parametric exponential family conditional distribution that resemble a truncated logistic regression model.

In future work, we plan to investigate the setting in which group membership attributes are not available at testing time. Extending our approach using a plug-in estimator of $P(a|\mathbf{x})$ in the fairness constraints introduces this estimator in the parametric form of the model. Understanding the impact of error from this estimator on predictor fairness in both theory and practice is an important direction of research.

Acknowledgments

This work was supported, in part, by the National Science Foundation under Grant No. 1652530 and by the Future of Life Institute (futureoflife.org) FLI-RFP-AI1 program.

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*.

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

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., and Keswani, V. 2019. Improved adversarial learning for fair classification. *arXiv preprint*.

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**.

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

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*.

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

Delage, E., and Ye, Y. 2010. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*.

Dheeru, D., and Karra Taniskidou, E. 2017. UCI machine learning repository.

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*.

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

Fathony, R.; Rezaei, A.; Bashiri, M.; Zhang, X.; and Ziebart, B. D. 2018b. Distributionally robust graphical models. In *NeurIPS*.

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*.

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

Hacker, P., and Wiedemann, E. 2017. A continuous framework for fairness. *arXiv preprint*.

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

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

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*.

Krasanakis, E.; Spyromitros-Xioufis, E.; Papadopoulos, S.; and Kompatsiaris, Y. 2018. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *WWW*.

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

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

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*.

Manning, C., and Klein, D. 2003. Optimization, maxent models, and conditional estimation without magic. In *NAACL*.

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.

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*.

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

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).

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

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

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*.

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*.

Zhang, L.; Wu, Y.; and Wu, X. 2018. Achieving non-discrimination in prediction. In *IJCAI*.