Characterizing Fairness Over the Set of Good Models Under Selective Labels

Amanda Coston [*1] **Ashesh Rambachan** [*2] **Alexandra Chouldechova** [3]

Abstract

Algorithmic risk assessments are used to inform decisions in a wide variety of high-stakes settings. Often multiple predictive models deliver similar overall performance but differ markedly in their predictions for individual cases, an empirical phenomenon known as the "Rashomon Effect." These models may have different properties over various groups, and therefore have different predictive fairness properties. We develop a framework for characterizing predictive fairness properties over the set of models that deliver similar overall performance, or "the set of good models." Our framework addresses the empirically relevant challenge of selectively labelled data in the setting where the selection decision and outcome are unconfounded given the observed data features. Our framework can be used to 1) audit for predictive bias; or 2) replace an existing model with one that has better fairness properties. We illustrate these use cases on a recidivism prediction task and a real-world credit-scoring task.

1. Introduction

Algorithmic risk assessments are used to inform decisions in high-stakes settings such as health care, child welfare, criminal justice, consumer lending and hiring (Caruana et al., 2015; Chouldechova et al., 2018; Kleinberg et al., 2018; Fuster et al., 2020; Raghavan et al., 2020). Unfettered use of such algorithms in these settings risks disproportionate harm to marginalized or protected groups (Barocas & Selbst, 2016; Dastin, 2018; Vigdor, 2019). As a result, there is widespread interest in measuring and limiting predictive disparities across groups.

The vast literature on algorithmic fairness offers numerous methods for learning anew the best performing model among those that satisfy a chosen notion of predictive fairness (e.g. Zemel et al. (2013), Agarwal et al. (2018), Agarwal et al. (2019)). However, for real-world settings where a risk assessment is already in use, practitioners and auditors may instead want to assess disparities with respect to the current model, which we term the *benchmark model*. For example, the benchmark model for a bank may be an existing credit score used to approve loans. The relevant question for practitioners is: Can we improve upon the benchmark model in terms of predictive fairness with minimal change in overall accuracy?

We explore this question through the lens of the "Rashomon Effect," a common empirical phenomenon whereby multiple models perform similarly overall but differ markedly in their predictions for individual cases (Breiman, 2001). These models may perform quite differently over various groups, and therefore have different predictive fairness properties. We propose an algorithm, Fairness in the Rashomon Set (FaiRS), to characterize predictive fairness properties over the set of models that perform similarly to a chosen benchmark model. We refer to this set as *the set of good models* (Dong & Rudin, 2020). FaiRS is designed to efficiently answer the following questions: What are the range of predictive disparities that could be generated over the set of good models? What is the disparity minimizing model within the set of good models?

A key empirical challenge in domains such as credit lending is that outcomes are not observed for all cases (Lakkaraju et al., 2017; Kleinberg et al., 2018). This *selective labels problem* is particularly vexing in the context of assessing predictive fairness. Our framework addresses selectively labelled data in contexts where the selection decision and outcome are unconfounded given the observed data features.

Our methods are useful for legal audits of disparate impact. In various domains, decisions that generate disparate impact must be justified by "business necessity" (civ, 1964; ECO, 1974; Barocas & Selbst, 2016). For instance, financial regulators investigate whether credit lenders could have offered more loans to minority applicants without affecting default rates (Gillis, 2019). Employment regulators may investigate whether resume screening software screens out underrepresented applicants for reasons that cannot be attributed to

[*]Equal contribution [1]Heinz College and Machine Learning Department, Carnegie Mellon University [2]Department of Economics, Harvard University [3]Heinz College, Carnegie Mellon Univesrity. Correspondence to: Amanda Coston <acoston@andrew.cmu.edu>, Ashesh Rambachan <asheshr@g.harvard.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

the job criteria (Raghavan et al., 2020). Our methods provide one possible formalization of the business necessity criteria. An auditor can use FaiRS to assess whether there exists an alternative model that reduces predictive disparities without compromising performance relative to the benchmark model. If possible, then it is difficult to justify the benchmark model on the grounds of business necessity.

Our methods can also be a useful tool for decision makers who want to improve upon an existing model. A decision maker may use FaiRS to search for a prediction function that reduces predictive disparities without compromising performance relative to the benchmark model. We emphasize that the effective usage of our methods requires careful thought about the broader social context surrounding the setting of interest (Selbst et al., 2019; Holstein et al., 2019).

Contributions: We (1) develop an algorithmic framework, Fairness in the Rashomon Set (FaiRS), to investigate predictive disparities over the set of good models; (2) provide theoretical guarantees on the generalization error and predictive disparities of FaiRS [§ 4]; (3) propose a variant of FaiRS that addresses the selective labels problem and achieves the same guarantees under oracle access to the outcome regression function [§ 5]; (4) use FaiRS to audit the COMPAS risk assessment, finding that it generates larger predictive disparities between black and white defendants than any model in the set of good models [§ 6]; and (5) use FaiRS on a selectively labelled credit-scoring dataset to build a model with lower predictive disparities than a benchmark model [§ 7]. All proofs are given in the Supplement.

2. Background and Related Work

2.1. Rashomon Effect

In a seminal paper on statistical modeling, Breiman (2001) observed that often a multiplicity of good models achieve similar accuracy by relying on different features, which he termed the "Rashomon effect." Even though they have similar accuracy, these models may differ along other key dimensions, and recent work considers the implications of the Rashomon effect for model simplicity, interpretability, and explainability (Fisher et al., 2019; Marx et al., 2020; Rudin, 2019; Dong & Rudin, 2020; Semenova et al., 2020).

We introduce these ideas into research on algorithmic fairness by studying the range of predictive disparities that can be achieved over the set of good models. We provide computational techniques to directly and efficiently investigate the range of predictive disparities that may be generated over the set of good models. Our recidivism risk prediction and credit scoring applications demonstrate that the set of good models is a rich empirical object, and we illustrate how characterizing the range of achievable predictive fairness properties over this set can be used for model learning and evaluation.

2.2. Fair Classification and Fair Regression

An influential literature on fair classification and fair regression constructs prediction functions that minimize loss subject to a predictive fairness constraint chosen by the decision maker (Dwork et al., 2012; Zemel et al., 2013; Hardt et al., 2016; Menon & Williamson, 2018; Donini et al., 2018; Agarwal et al., 2018; 2019; Zafar et al., 2019). In contrast, we construct prediction functions that minimize a chosen measure of predictive disparities subject to a constraint on overall performance. This is useful when decision makers find it difficult to specify acceptable levels of predictive disparities, but instead know what performance loss is tolerable. It may be unclear, for instance, how a lending institution should specify acceptable differences in credit risk scores across groups, but the lending institution can easily specify an acceptable average default rate among approved loans. Our methods allow users to directly search for prediction functions that reduce disparities given such a specified loss tolerance. Similar in spirit to our work, Zafar et al. (2019) provide a method for selecting a classifier that minimizes a particular notion of predictive fairness, "decision boundary covariance," subject to a performance constraint. Our method applies more generally to a large class of predictive disparities and covers both classification and regression tasks.

While originally developed to solve fair classification and fair regression problems, we show that the "reductions approach" used in Agarwal et al. (2018; 2019) can be suitably adapted to solve general optimization problems over the set of good models. This provides a general computational approach that may be useful for investigating the implications of the Rashomon Effect for other model properties.

In constructing the set of good models with comparable performance to a benchmark model, our work bears resemblance to techniques that "post-process" existing models. Post-processing techniques typically modify the predictions from an existing model to achieve a target notion of fairness (Hardt et al., 2016; Pleiss et al., 2017; Kim et al., 2019). By contrast, our methods only use the existing model to calibrate the performance constraint, but need not share any other properties with the benchmark model. While post-processing techniques often require access to individual predictions from the benchmark model, our approach only requires that we know its average loss.

2.3. Selective Labels and Missing Data

In settings such as criminal justice and credit lending, the training data only contain labeled outcomes for a selectively observed sample from the full population of interest. For example, banks use risk scores to assess all loan applicants,

yet the historical data only contains default/repayment outcomes for those applicants whose loans were approved. This is a missing data problem (Little & Rubin, 2019). Because the outcome label is missing based on a selection mechanism, this type of missing data is known as the *selective labels problem* (Lakkaraju et al., 2017; Kleinberg et al., 2018). One solution treats the selectively labelled population as if it were the population of interest, and proceeds with training and evaluation on the selectively labelled population only. This is also called the "*known good-bad*" (KGB) approach (Zeng & Zhao, 2014; Nguyen et al., 2016). However, evaluating a model on a population different than the one on which it will be used can be highly misleading, particularly with regards to predictive fairness measures (Kallus & Zhou, 2018; Coston et al., 2020). Unfortunately, most fair classification and fair regression methods do not offer modifications to address the selective labels problem. Our framework does [§ 5].

Popular in credit lending applications, "reject inference" procedures incorporate information from the selectively unobserved cases (i.e., rejected applicants) in model construction and evaluation by imputing missing outcomes using augmentation, reweighing or extrapolation-based approaches (Li et al., 2020; Mancisidor et al., 2020). These approaches are similar to domain adaptation techniques, and indeed the selective labels problem can be cast as domain adaptation since the labelled training data is not a random sample of the target distribution. Most relevant to our setting are covariate shift methods for domain adaptation. Reweighing procedures have been proposed for jointly addressing covariate shift and fairness (Coston et al., 2019; Singh et al., 2021). While FaiRS similarly uses iterative reweighing to solve our joint optimization problem, we explicitly use extrapolation to address covariate shift. Empirically we find extrapolation can achieve lower disparities than reweighing.

3. Setting and Problem Formulation

The population of interest is described by the random vector $(X_i, A_i, D_i, Y_i^*) \sim P$, where $X_i \in \mathcal{X}$ is a feature vector, $A_i \in \{0,1\}$ is a protected or sensitive attribute, $D_i \in \mathcal{D}$ is the decision and $Y_i^* \in \mathcal{Y} \subseteq [0,1]$ is a discrete or continuous outcome. The training data consist of n i.i.d. draws from the joint distribution P and may suffer from a *selective labels problem*: There exists $\mathcal{D}^* \subseteq \mathcal{D}$ such that the outcome is observed if and only if the decision satisfies $D_i \in \mathcal{D}^*$. Hence, the training data are $\{(X_i, A_i, D_i, Y_i)\}_{i=1}^n$, where $Y_i = Y_i^* 1\{D_i \in \mathcal{D}^*\})$ is the *observed outcome* and $1\{\cdot\}$ denotes the indicator function.

Given a specified set of prediction functions $\mathcal{F}$ with elements $f\colon \mathcal{X} \to [0,1]$, we search for the prediction function $f \in \mathcal{F}$ that minimizes or maximizes a measure of predictive disparities with respect to the sensitive attribute subject to a constraint on predictive performance. We measure performance using average loss, where $l\colon \mathcal{Y} \times [0,1] \to [0,1]$ is the loss function and $\text{loss}(f) := \mathbb{E}\left[l(Y_i^*, f(X_i))\right]$. The loss function is assumed to be 1-Lipshitz under the l_1-norm following Agarwal et al. (2019). The constraint on performance takes the form $\text{loss}(f) \leq \epsilon$ for some specified *loss tolerance* $\epsilon \geq 0$. The set of prediction functions satisfying this constraint is the *set of good models*.

The loss tolerance may be chosen based on an existing benchmark model $\tilde{f}$ such as an existing risk score, e.g., by setting $\epsilon = (1+\delta)\,\text{loss}(\tilde{f})$ for some $\delta \in [0,1]$. The set of good models now describes the set of models whose performance lies within a δ-neighborhood of the benchmark model. When defined in this manner, the set of good models is also called the "Rashomon set" (Rudin, 2019; Fisher et al., 2019; Dong & Rudin, 2020; Semenova et al., 2020).

3.1. Measures of Predictive Disparities

We consider measures of predictive disparity of the form

$$\text{disp}(f) := \beta_0 \mathbb{E}\left[f(X_i)|\mathcal{E}_{i,0}\right] + \beta_1 \mathbb{E}\left[f(X_i)|\mathcal{E}_{i,1}\right], \quad (1)$$

where $\mathcal{E}_{i,a}$ is a group-specific conditioning event that depends on (A_i, Y_i^*) and $\beta_a \in \mathbb{R}$ for $a \in \{0,1\}$ are chosen parameters. Note that we measure predictive disparities over the *full* population (i.e., not conditional on D_i).

For different choices of the conditioning events $\mathcal{E}_{i,0}, \mathcal{E}_{i,1}$ and parameters β_0, β_1, our predictive disparity measure summarizes violations of common definitions of predictive fairness.

Definition 1. ***Statistical parity*** *(SP) requires the prediction $f(X_i)$ to be independent of the attribute A_i (Dwork et al., 2012; Zemel et al., 2013; Feldman et al., 2015). By setting $\mathcal{E}_{i,a} = \{A_i = a\}$ for $a \in \{0,1\}$ and $\beta_0 = -1$, $\beta_1 = 1$, $\text{disp}(f)$ measures the difference in average predictions across values of the sensitive attribute.*

Definition 2. *Suppose $\mathcal{Y} = \{0,1\}$.* ***Balance for the positive class*** *(BFPC) and* ***balance for the negative class*** *(BFNC) requires the prediction $f(X_i)$ to be independent of the attribute A_i conditional on $Y_i^* = 1$ and $Y_i^* = 0$ respectively (e.g., Chapter 2 of (Barocas et al., 2019)). Defining $\mathcal{E}_{i,a} = \{Y_i^* = 1, A_i = a\}$ for $a \in \{0,1\}$ and $\beta_0 = -1, \beta_1 = 1$, $\text{disp}(f)$ describes the difference in average predictions across values of the sensitive attribute given $Y_i^* = 1$. If instead $\mathcal{E}_{i,a} = \{Y_i^* = 0, A_i = a\}$ for $a \in \{0,1\}$, then $\text{disp}(f)$ equals the difference in average predictions across values of the sensitive attribute given $Y_i^* = 0$.*

Our focus on differences in average predictions across groups is a common relaxation of parity-based predictive fairness definitions (Corbett-Davies et al., 2017; Mitchell et al., 2019).

Our predictive disparity measure can also be used for *fairness promoting interventions*, which aim to increase opportunities for a particular group. For instance, the decision maker may wish to search for the prediction function among the set of good models that minimizes the average predicted risk score $f(X_i)$ for a historically disadvantaged group.

Definition 3. *Defining $\mathcal{E}_{i,1} = \{A_i = 1\}$ and $\beta_0 = 0, \beta_1 = 1$, $\text{disp}(f)$ measures the average risk score for the group with $A_i = 1$. This is an **affirmative action**-based fairness promoting intervention. Further assuming $\mathcal{Y} = \{0, 1\}$ and defining $\mathcal{E}_{i,1} = \{Y_i^* = 1, A_i = 1\}$, $\text{disp}(f)$ measures the average risk score for the group with both $Y_i^* = 1$, $A_i = 1$. This is a **qualified affirmative action**-based fairness promoting intervention.*

Our approach can accommodate other notions of predictive disparities. In Supplement §A.4, we show how to achieve *bounded group loss*, which requires that the average loss conditional on each value of the sensitive attribute reach some threshold (Agarwal et al., 2019).

3.2. Characterizing Predictive Disparities over the Set of Good Models

We develop the algorithmic framework, Fairness in the Rashomon Set (FaiRS), to solve two related problems over the set of good models. First, we characterize the range of predictive disparities by minimizing or maximizing the predictive disparity measure over the set of good models. We focus on the minimization problem

$$\min_{f \in \mathcal{F}} \text{disp}(f) \text{ s.t. } \text{loss}(f) \leq \epsilon. \tag{2}$$

Second, we search for the prediction function that minimizes the absolute predictive disparity over the set of good models

$$\min_{f \in \mathcal{F}} |\text{disp}(f)| \text{ s.t. } \text{loss}(f) \leq \epsilon. \tag{3}$$

For auditors, (2) traces out the range of predictive disparities that *could* be generated in a given setting, thereby identifying where the benchmark model lies on this frontier. This is crucially related to the legal notion of "business necessity" in assessing disparate impact – the regulator may audit whether there exist alternative prediction functions that achieve similar performance yet generate different predictive disparities (civ, 1964; ECO, 1974; Barocas & Selbst, 2016). For decision makers, (3) searches for prediction functions that reduce absolute predictive disparities without compromising predictive performance.

4. A Reductions Approach to Optimizing over the Set of Good Models

We characterize the range of predictive disparities (2) and find the absolute predictive disparity minimizing model (3) over the set of good models using techniques inspired by the reductions approach in Agarwal et al. (2018; 2019). Although originally developed to solve fair classification and fair regression problems in the case without selective labels, we extend the reductions approach to solve general optimization problems over the set of good models in the presence of selective labels. For exposition, we first focus on the case without selective labels, where $\mathcal{D}^* = \mathcal{D}$ and the outcome Y_i^* is observed for all observations. We solve (2) in the main text and (3) in § A.3 of the Supplement. We cover selective labels in § 5.

4.1. Computing the Range of Predictive Disparities

We consider randomized prediction functions that select $f \in \mathcal{F}$ according to some distribution $Q \in \Delta(\mathcal{F})$ where Δ denotes the probability simplex. Let $\text{loss}(Q) := \sum_{f \in \mathcal{F}} Q(f) \text{loss}(f)$ and $\text{disp}(Q) := \sum_{f \in \mathcal{F}} Q(f) \text{disp}(f)$. We solve

$$\min_{Q \in \Delta(\mathcal{F})} \text{disp}(Q) \text{ s.t. } \text{loss}(Q) \leq \epsilon. \tag{4}$$

While it may be possible to solve this problem directly for certain parametric function classes, we develop an approach that can be applied to any generic function class.[1] A key object for doing so will be classifiers obtained by thresholding prediction functions. For cutoff $z \in [0, 1]$, define $h_f(x, z) = 1\{f(x) \geq z\}$ and let $\mathcal{H} := \{h_f : f \in \mathcal{F}\}$ be the set of all classifiers obtained by thresholding prediction functions $f \in \mathcal{F}$. We first reduce the optimization problem (4) to a constrained classification problem through a discretization argument, and then solve the resulting constrained classification problem through a further reduction to finding the saddle point of a min-max problem.

Following the notation in Agarwal et al. (2019), we define a discretization grid for $[0, 1]$ of size N with $\alpha := 1/N$ and $\mathcal{Z}_\alpha := \{j\alpha \colon j = 1, \ldots, N\}$. Let $\tilde{\mathcal{Y}}_\alpha$ be an $\frac{\alpha}{2}$-cover of $\mathcal{Y}$. The piecewise approximation to the loss function is $l_\alpha(y, u) := l(\underline{y}, [u]_\alpha + \frac{\alpha}{2})$, where $\underline{y}$ is the smallest $\tilde{y} \in \tilde{\mathcal{Y}}_\alpha$ such that $|y - \tilde{y}| \leq \frac{\alpha}{2}$ and $[u]_\alpha$ rounds u down to the nearest integer multiple of α. For a fine enough discretization grid, $\text{loss}_\alpha(f) := \mathbb{E}\left[l_\alpha(Y_i^*, f(X_i))\right]$ approximates $\text{loss}(f)$.

Define $c(y, z) := N \times \left(l(y, z + \frac{\alpha}{2}) - l(y, z - \frac{\alpha}{2})\right)$ and Z_α to be the random variable that uniformly samples $z_\alpha \in \mathcal{Z}_\alpha$ and is independent of the data (X_i, A_i, Y_i^*). For $h_f \in \mathcal{H}$, define the cost-sensitive average loss function as $\text{cost}(h_f) := \mathbb{E}\left[c(\underline{Y_i^*}, Z_\alpha) h_f(X_i, Z_\alpha)\right]$. Lemma 1 in Agarwal et al. (2019) shows $\text{cost}(h_f) + c_0 = \text{loss}_\alpha(f)$ for any $f \in \mathcal{F}$, where $c_0 \geq 0$ is a constant that does not depend on f. Since $\text{loss}_\alpha(f)$ approximates $\text{loss}(f)$, $\text{cost}(h_f)$ also approximates $\text{loss}(f)$. For $Q \in \Delta(\mathcal{F})$, define $Q_h \in \Delta(\mathcal{H})$

[1] Our error analysis only covers function classes whose Rademacher complexity can be bounded as in Assumption 1.

to be the induced distribution over threshold classifiers h_f. By the same argument, $\text{cost}(Q_h) + c_0 = \text{loss}_\alpha(Q)$, where $\text{cost}(Q_h) := \sum_{h_f \in \mathcal{H}} Q_h(h)\,\text{cost}(h_f)$ and $\text{loss}_\alpha(Q)$ is defined analogously.

We next relate the predictive disparity measure defined on prediction functions to a predictive disparity measure defined on threshold classifiers. Define $\text{disp}(h_f) := \beta_0 \mathbb{E}\left[h_f(X_i, Z_\alpha) \mid \mathcal{E}_{i,0}\right] + \beta_1 \mathbb{E}\left[h_f(X_i, Z_\alpha) \mid \mathcal{E}_{i,1}\right]$.

Lemma 1. *Given any distribution over (X_i, A_i, Y_i^*) and $f \in \mathcal{F}$, $|\text{disp}(h_f) - \text{disp}(f)| \leq (|\beta_0| + |\beta_1|)\,\alpha$.*

Lemma 1 combined with Jensen's Inequality imply $|\,\text{disp}(Q_h) - \text{disp}(Q)| \leq (|\beta_0| + |\beta_1|)\,\alpha$.

Based on these results, we approximate (4) with its analogue over threshold classifiers

$$\min_{Q_h \in \Delta(\mathcal{H})} \text{disp}(Q_h) \text{ s.t. } \text{cost}(Q_h) \leq \epsilon - c_0. \tag{5}$$

We solve the sample analogue in which we minimize $\widehat{\text{disp}}(Q_h)$ subject to $\widehat{\text{cost}}(Q_h) \leq \hat{\epsilon}$, where $\hat{\epsilon} := \epsilon - \hat{c}_0$ plus additional slack, and $\hat{c}_0, \widehat{\text{disp}}(Q_h), \widehat{\text{cost}}(Q_h)$ are the associated sample analogues. We form the Lagrangian $L(Q_h, \lambda) := \widehat{\text{disp}}(Q_h) + \lambda(\widehat{\text{cost}}(Q_h) - \hat{\epsilon})$ with primal variable $Q_h \in \Delta(\mathcal{H})$ and dual variable $\lambda \in \mathbb{R}^+$. Solving the sample analogue is equivalent to finding the saddle point of the min-max problem $\min_{Q_h \in \Delta(\mathcal{H})} \max_{0 \leq \lambda \leq B_\lambda} L(Q_h, \lambda)$, where $B_\lambda \geq 0$ bounds the Lagrange multiplier. We search for the saddle point by adapting the exponentiated gradient algorithm used in Agarwal et al. (2018; 2019). The algorithm delivers a ν-approximate saddle point of the Lagrangian, denoted $(\hat{Q}_h, \hat{\lambda})$. Since it is standard, we provide the details of and the pseudocode for the exponentiated gradient algorithm in § A.1 of the Supplement.

4.2. Error Analysis

The suboptimality of the returned solution $\hat{Q}_h$ can be controlled under conditions on the complexity of the model class $\mathcal{F}$ and how various parameters are set.

Assumption 1. *Let $R_n(\mathcal{H})$ be the Rademacher complexity of $\mathcal{H}$. There exists constants $C, C', C'' > 0$ and $\phi \leq 1/2$ such that $R_n(\mathcal{H}) \leq Cn^{-\phi}$ and $\hat{\epsilon} = \epsilon - \hat{c}_0 + C'n^{-\phi} - C''n^{-1/2}$.*

Theorem 1. *Suppose Assumption 1 holds for $C' \geq 2C + 2 + \sqrt{2\ln(8N/\delta)}$ and $C'' \geq \sqrt{\frac{-\log(\delta/8)}{2}}$. Let n_0, n_1 denote the number of samples satisfying the events $\mathcal{E}_{i,0}, \mathcal{E}_{i,1}$ respectively.*

Then, the exponentiated gradient algorithm with $\nu \propto n^{-\phi}$, $B_\lambda \propto n^\phi$ and $N \propto n^\phi$ terminates in $O(n^{4\phi})$ iterations and returns $\hat{Q}_h$, which when viewed as a distribution over $\mathcal{F}$, satisfies with probability at least $1 - \delta$ one of the following: 1) $\hat{Q}_h \neq$ null, $\text{loss}(\hat{Q}_h) \leq \epsilon + \tilde{O}(n^{-\phi})$ and $\text{disp}(\hat{Q}_h) \leq \text{disp}(\tilde{Q}) + \tilde{O}(n_0^{-\phi}) + \tilde{O}(n_1^{-\phi})$ for any $\tilde{Q}$ that is feasible in (4); or 2) $\hat{Q}_h =$ null and (4) is infeasible.[2]

Theorem 1 shows that the returned solution $\hat{Q}_h$ is approximately feasible and achieves the lowest possible predictive disparity up to some error. Infeasibility is a concern if no prediction function $f \in \mathcal{F}$ satisfies the average loss constraint. Assumption 1 is satisfied for instance under LASSO and ridge regression. If Assumption 1 does not hold, FaiRS still delivers good solutions to the sample analogue of Eq. 5 (see Supplement § C.1.2).

A practical challenge is that the solution returned by the exponentiated gradient algorithm $\hat{Q}_h$ is a stochastic prediction function with possibly large support. Therefore it may be difficult to describe, time-intensive to evaluate, and memory-intensive to store. Results from Cotter et al. (2019) show that the support of the returned stochastic prediction function may be shrunk while maintaining the same guarantees on its performance by solving a simple linear program. The linear programming reduction reduces the stochastic prediction function to have at most two support points and we use this linear programming reduction in our empirical work (see § A.2 of the Supplement for details).

5. Optimizing Over the Set of Good Models Under Selective Labels

We now modify the reductions approach to the empirically relevant case in which the training data suffer from the selective labels problem, whereby the outcome Y_i^* is observed only if $D_i \in \mathcal{D}^*$ with $\mathcal{D}^* \subset \mathcal{D}$. The main challenge concerns evaluating model properties over the target population when we only observe labels for a selective (i.e., biased) sample. We propose a solution that uses outcome modeling, also known as extrapolation, to estimate these properties.

To motivate this approach, we observe that average loss and measures of predictive disparity (1) that condition on Y_i^* are not identified under selective labels without further assumptions. We introduce the following assumption on the nature of the selective labels problem for the binary decision setting with $\mathcal{D} = \{0, 1\}$ and $\mathcal{D}^* = \{1\}$.

Assumption 2. *The joint distribution $(X_i, A_i, D_i, Y_i^*) \sim P$ satisfies 1) **selection on observables**: $D_i \perp\!\!\!\perp Y_i^* \mid X_i$, and 2) **positivity**: $\mathbb{P}(D_i = 1 \mid X_i = x) > 1$ with probability one.*

This assumption is common in causal inference and selection bias settings (e.g., Chapter 12 of Imbens & Rubin (2015) and Heckman (1990))[3] and in covariate shift learning (Moreno-Torres et al., 2012). Under Assumption 2, the

[2] The notation $\tilde{O}(\cdot)$ suppresses polynomial dependence on $\ln(n)$ and $\ln(1/\delta)$

[3] Casting this into potential outcomes notation where Y_i^d is the

regression function $\mu(x) := \mathbb{E}[Y_i^* \mid X_i = x]$ is identified as $\mathbb{E}[Y_i \mid X_i, D_i = 1]$, and may be estimated by regressing the observed outcome Y_i on the features X_i among observations with $D_i = 1$, yielding the outcome model $\hat{\mu}(x)$.

We can use the outcome model to estimate loss on the full population. One approach, *Reject inference by extrapolation* (RIE), uses $\hat{\mu}(x)$ as pseudo-outcomes for the unknown observations (Crook & Banasik, 2004). We consider a second approach, *Interpolation & extrapolation* (IE), which uses $\hat{\mu}(x)$ as pseudo-outcomes for *all* applicants, replacing the $\{0, 1\}$ labels for known cases with smoothed estimates of their underlying risks. Letting n^0, n^1 be the number of observations in the training data with $D_i = 0$, $D_i = 1$ respectively, Algorithms 1-2 summarize the RIE and IE methods. If the outcome model could perfectly recover $\mu(x)$, then the IE approach recovers an oracle setting for which the FaiRS error analysis continues to hold (Theorem 2 below).

Algorithm 1: Reject inference by extrapolation (RIE) for the selective labels setting

Input: $\{(X_i, Y_i, D_i = 1, A_i)\}_{i=1}^{n^1}$, $\{(X_i, D_i = 0, A_i)\}_{i=1}^{n^0}$
Estimate $\hat{\mu}(x)$ by regressing $Y_i \sim X_i \mid D_i = 1$.
$\hat{Y}(X_i) \leftarrow (1 - D_i)\hat{\mu}(X_i) + D_i Y_i$
Output: $\{(X_i, \hat{Y}_i(X_i), D_i, A_i)\}_{i=1}^{n^1}$, $\{(X_i, \hat{Y}_i(X_i), D_i, A_i)\}_{i=1}^{n^0}$

Algorithm 2: Interpolation and extrapolation (IE) method for the selective labels setting

Input: $\{(X_i, Y_i, D_i = 1, A_i)\}_{i=1}^{n^1}$, $\{(X_i, D_i = 0, A_i)\}_{i=1}^{n^0}$
Estimate $\hat{\mu}(x)$ by regressing $Y_i \sim X_i \mid D_i = 1$.
$\hat{Y}(X_i) \leftarrow \hat{\mu}(X_i)$
Output: $\{(X_i, \hat{Y}_i(X_i), D_i, A_i)\}_{i=1}^{n^1}$, $\{(X_i, \hat{Y}_i(X_i), D_i, A_i)\}_{i=1}^{n^0}$

Estimating predictive disparity measures on the full population requires a more general definition of predictive disparity than previously given in Eq. 1. Define the modified predictive disparity measure over threshold classifiers as

$$\text{disp}(h_f) = \beta_0 \frac{\mathbb{E}\left[g(X_i, Y_i) h_f(X_i, Z_\alpha) \mid \mathcal{E}_{i,0}\right]}{\mathbb{E}[g(X_i, Y_i) \mid \mathcal{E}_{i,0}]} + \beta_1 \frac{\mathbb{E}\left[g(X_i, Y_i) h_f(X_i, Z_\alpha) | \mathcal{E}_{i,1}\right]}{\mathbb{E}[g(X_i, Y_i) \mid \mathcal{E}_{i,1}]}, \quad (6)$$

where the nuisance function $g(X_i, Y_i)$ is constructed to identify the measure of interest.[4] To illustrate, the qualified affirmative action fairness-promoting intervention (Def. 3) is identified as $\mathbb{E}[f(X_i)|Y_i^* = 1, A_i = 1] = \frac{\mathbb{E}[f(X_i)\mu(X_i)|A_i=1]}{\mathbb{E}[\mu(X_i)|A_i=1]}$ under Assumption 2 (See proof of Lemma 8 in the Supplement). This may be estimated by plugging in the outcome model estimate $\hat{\mu}(x)$. Therefore, Eq. 6 specifies the qualified affirmative action fairness-promoting intervention by setting $\beta_0 = 0$, $\beta_1 = 1$, $\mathcal{E}_{i,1} = 1\{A_i = 1\}$, and $g(X_i, Y_i) = \hat{\mu}(X_i)$. This more general definition (Eq. 6) is only required for predictive disparity measures that condition on events $\mathcal{E}$ depending on both Y^* and A; it is straightforward to compute disparities based on events $\mathcal{E}$ that only depend on A over the full population. To compute disparities based on events $\mathcal{E}$ that also depend on Y^*, we find the saddle point of the following Lagrangian: $L(h_f, \lambda) = \hat{\mathbb{E}}\left[\mathbb{E}_{Z_\alpha}\left[c_\lambda(\underline{\hat{\mu}}_i, A_i, Z_\alpha) h_f(X_i, Z_\alpha)\right]\right] - \lambda\hat{\epsilon}$, where we now use case weights $c_\lambda(\underline{\hat{\mu}}_i, A_i, Z_\alpha) := \frac{\beta_0}{\hat{p}_0} g(X_i, Y_i)(1 - A_i) + \frac{\beta_1}{\hat{p}_1} g(X_i, Y_i) A_i + \lambda c(\underline{\hat{\mu}}_i, Z_\alpha)$ with $\hat{p}_a = \hat{\mathbb{E}}[g(X_i, Y_i) 1\{A_i = a\}]$ for $a \in \{0, 1\}$. Finally, as before, we find the saddle point using the exponentiated gradient algorithm.

5.1. Error Analysis under Selective Labels

Define $\text{loss}_\mu(f) := \mathbb{E}[l(\mu(X_i), f(X_i))]$ for $f \in \mathcal{F}$ with $\text{loss}_\mu(Q)$ defined analogously for $Q \in \Delta(\mathcal{F})$. The error analysis of the exponentiated gradient algorithm continues to hold in the presence of selective labels under oracle access to the true outcome regression function μ.

Theorem 2 (Selective Labels)**.** *Suppose Assumption 2 holds and the exponentiated gradient algorithm is given as input the modified training data* $\{(X_i, A_i, \mu(X_i)\}_{i=1}^n$.

Under the same conditions as Theorem 1, the exponentiated gradient algorithm terminates in $O(n^{4\phi})$ *iterations and returns* $\hat{Q}_h$*, which when viewed as a distribution over* $\mathcal{F}$*, satisfies with probability at least* $1 - \delta$ *either one of the following: 1)* $\hat{Q}_h \neq null$*,* $\text{loss}_\mu(\hat{Q}_h) \leq \epsilon + \tilde{O}(n^{-\phi})$ *and* $\text{disp}(\hat{Q}_h) \leq \text{disp}(\tilde{Q}) + \tilde{O}(n_0^{-\phi}) + \tilde{O}(n_1^{-\phi})$ *for any* $\tilde{Q}$ *that is feasible in (4); or 2)* $\hat{Q}_h = null$ *and (4) is infeasible.*

In practice, estimation error in $\hat{\mu}$ will affect the bounds in Theorem 2. The empirical analysis in § 7 finds that our method nonetheless performs well when using $\hat{\mu}$.

6. Application: Recidivism Risk Prediction

We use FaiRS to empirically characterize the range of disparities over the set of good models in a recidivism risk pre-

counterfactual outcome if decision d were assigned, we define $Y_i^0 = 0$ and $Y_i^1 = Y_i^*$ (e.g., a rejected loan application cannot default). The observed outcome Y_i then equals $Y_i^1 D_i$.

[4] Note that we state this general form of g to allow g to use Y_i for e.g. doubly-robust style estimates.

diction task applied to ProPublica's COMPAS data (Angwin et al., 2016). Our goal is to illustrate (i) how FaiRS may be used to tractably characterize the range of predictive disparities over the set of good models; (ii) that the range of predictive disparities over the set of good models can be quite large empirically; and (iii) how an auditor may use the set of good models to assess whether the COMPAS risk assessment generates larger disparities than other competing good models. Such an analysis is a crucial step to assessing legal claims of disparate impact.

COMPAS is a proprietary risk assessment developed by Northpointe (now Equivant) using up to 137 features (Rudin et al., 2020). As this data is not publicly available, our audit makes use of ProPublica's COMPAS dataset which contains demographic information and prior criminal history for criminal defendants in Broward County, Florida. Lacking access to the data used to train COMPAS, our set of good models may not include COMPAS itself (Angwin et al., 2016). Nonetheless, prior work has shown that simple models using age and criminal history perform on par with COMPAS (Angelino et al., 2018). These features will therefore suffice to perform our audit. A notable limitation of the ProPublica COMPAS dataset is that it does not contain information for defendants who remained incarcerated. Lacking both features and outcomes for this group, we proceed without addressing this source of selection bias. We also make no distinction between criminal defendants who had varying lengths of incarceration before release, effectively assuming a null treatment effect of incarceration on recidivism. This assumption is based on findings that a counterfactual audit of COMPAS yields equivalent conclusions (Mishler, 2019).

We analyze the range of predictive disparities with respect to race for three common notions of fairness (Definitions 1-2) among logistic regression models on a quadratic polynomial of the defendant's age and number of prior offenses whose training loss is near-comparable to COMPAS (loss tolerance $\epsilon = 1\%$ of COMPAS training loss).[5] We split the data 50%-50% into a train and test set. Table 1 summarizes the range of predictive disparities on the test set. The disparity minimizing and disparity maximizing models over the set of good of models achieve a test loss that is comparable to COMPAS (see § D.1 of the Supplement).

For each predictive disparity measure, the set of good models includes models that achieve significantly lower disparities than COMPAS. In this sense, COMPAS generates "unjustified" disparate impact as there exists competing models that would reduce disparities without compromising performance. Notably, COMPAS' disparities are also larger than the maximum disparity over the set of good models. For example, the difference in COMPAS' average predictions for black relative to white defendants is strictly larger than that of any model in the set of good models (Table 1, SP). Interestingly, the minimal balance for the positive class and balance for the negative class disparities between black and white defendants over the set of good models are strictly positive (Table 1, BFPC and BFNC). For example any model whose performance lies in a neighborhood of COMPAS' loss has a higher false positive rate for black defendants than white defendants. This suggests while we can reduce predictive disparities between black and white defendants relative to COMPAS on all measures, we may be unable to eliminate balance for the positive class and balance for the negative class disparities without harming predictive performance.

[5] We use a quadratic form following the analysis in Rudin et al. (2020).

Table 1. COMPAS fails an audit of the "business necessity" defense for disparate impact by race. The set of good models (performing within 1% of COMPAS's training loss) includes models that achieve significantly lower disparities than COMPAS. The first panel (SP) displays the disparity in average predictions for black versus white defendants (Def. 1). The second panel (BFPC) analyzes the disparity in average predictions for black versus white defendants in the positive class, and the third panel examines the disparity in average predictions for black versus white defendants in the negative class (Def. 2). Standard errors are reported in parentheses. See § 6 for details.

	MIN. DISP.	MAX. DISP.	COMPAS
SP	−0.060 (0.004)	0.120 (0.007)	0.194 (0.013)
BFPC	0.049 (0.005)	0.125 (0.012)	0.156 (0.016)
BFNC	0.044 (0.005)	0.117 (0.009)	0.174 (0.016)

In addition to the retrospective auditing considered in this section, characterizing the range of predictive disparities over the set of good models is also important for model development and selection. The next section shows how to construct a more equitable model that performs comparably to a benchmark.

7. Application: Consumer Lending

Suppose a financial institution wishes to replace an existing credit scoring model with one that has better fairness properties and comparable performance, if such a model exists. We show how to accomplish this task by using FaiRS to find the absolute predictive disparity-minimizing model over the set of good models. On a real world consumer lending dataset with selectively labeled outcomes, we find that this approach yields a model that reduces predictive disparities relative to the benchmark without compromising overall performance.

We use data from Commonwealth Bank of Australia, a large financial institution in Australia (henceforth, "CommBank"), on a sample of 7,414 personal loan applications submitted from July 2017 to July 2019 by customers that did not have a prior financial relationship with CommBank. A personal loan is a credit product that is paid back with monthly installments and used for a variety of purposes such as purchasing a used car or refinancing existing debt. In our sample, the median personal loan size is AU$10,000 and the median interest rate is 13.9% per annum. For each loan application, we observe application-level information such as the applicant's credit score and reported income, whether the application was approved by CommBank, the offered terms of the loan, and whether the applicant defaulted on the loan. There is a selective labels problem as we only observe whether an applicant defaulted on the loan within 5 months (Y_i) if the application was funded, where "funded" denotes that the application is both approved by CommBank and the offered terms were accepted by the applicant. In our sample, 44.9% of applications were funded and 2.0% of funded loans defaulted within 5 months.

Motivated by a decision maker that wishes to reduce credit access disparities across geographic regions, we focus on the task of predicting the likelihood of default $Y_i^* = 1$ based on information in the loan application X_i while limiting predictive disparities across SA4 geographic regions within Australia. SA4 regions are statistical geographic areas defined by the Australian Bureau of Statistics (ABS) and are analogous to counties in the United States. An SA4 region is classified as socioeconomically disadvantaged ($A_i = 1$) if it falls in the top quartile of SA4 regions based on the ABS' Index of Relative Socioeconomic Disadvantage (IRSD), which is an index that aggregates census data related to socioeconomic disadvantage.[6] Applicants from disadvantaged SA4 regions are under-represented among funded applications, comprising 21.7% of all loan applications, but only 19.7% of all funded loan applications.

Our experiment investigates the performance of FaiRS under our two proposed extrapolation-based solutions to selective labels, RIE and IE (See Algorithms 1-2), as well as the Known-Good Bad (KGB) approach that uses only the selectively labelled population. Because we do not observe default outcomes for all applications, we conduct a semi-synthetic simulation experiment by generating synthetic funding decisions and default outcomes. On a 20% sample of applicants, we learn $\pi(x) := \hat{P}(D_i = 1|X_i = x)$ and $\mu(x) := \hat{P}(Y_i = 1|X_i = x, D_i = 1)$ using random forests. We generate synthetic funding decisions $\widetilde{D}_i$ according to $\widetilde{D}_i \mid X_i \sim Bernoulli(\pi(X_i))$ and synthetic default outcomes $\widetilde{Y}_i^*$ according to $\widetilde{Y}_i^* \mid X_i \sim Bernoulli(\mu(X_i))$. We train all models as if we only knew the synthetic outcome for the synthetically funded applications. We estimate $\hat{\mu}(x) := \hat{P}(\tilde{Y}_i = 1|X_i = x, \tilde{D}_i = 1)$ using random forests and use $\hat{\mu}(X_i)$ to generate the pseudo-outcomes $\hat{Y}(X_i)$ for RIE and IE as described in Algorithms 1 and 2. As benchmark models, we use the loss-minimizing linear models learned using KGB, RIE, and IE approaches, whose respective training losses are used to select the corresponding loss tolerances ϵ. We use the class of linear models for the FaiRS algorithm for KGB, RIE, and IE approaches.

We compare against the fair reductions approach to classification (fairlearn) and the Target-Fair Covariate Shift (TFCS) method. TFCS iteratively reweighs the training data via gradient descent on an objective function comprised of the covariate shift-reweighed classification loss and a fairness loss (Coston et al., 2019). Fairlearn searches for the loss-minimizing model subject to a fairness parity constraint (Agarwal et al., 2018). The fairlearn model is effectively a KGB model since the fairlearn package does not offer modifications for selective labels.[7] We use logistic regression as the base model for both fairlearn and TFCS. Results are reported on all applicants in a held out test set, and performance metrics are constructed with respect to the synthetic outcome $\tilde{Y}_i^*$.

Figure 1 shows the AUC (y-axis) against disparity (x-axis) for the KGB, RIE, IE benchmarks and their FaiRS variants as well as the TFCS models and fairlearn models. Colors denote the adjustment strategy for selective labels, and the shape specifies the optimization method. The first row evaluates the models on all applicants in the test set (i.e., the target population). On the target population, FaiRS with reject extrapolation (RIE and IE) reduces disparities while achieving performance comparable to the benchmarks and to the reweighing approach (TFCS). It also achieves lower disparities than TFCS, likely because TFCS optimizes a non-convex objective function and may therefore converge to a local minimum. Reject extrapolation achieves better AUC than all KGB models, and only one KGB model (fairlearn) achieves a lower disparity. The second row evaluates the models on only the *funded* applicants. Evaluation on the funded cases underestimates disparities across the methods and overestimates AUC for the TFCS and KGB models. This underscores the importance of accounting for the selective labels problem in both model construction and evaluation.

FaiRS is also applicable in the regression setting. On the

[6] Complete details on the IRSD may be found in Australian Bureau of Statistics (2016) and additional details on the definition of socioeconomic disadvantage are given in § D.2 of the Supplement.

[7] To accommodate reject inference, a method must support real-valued outcomes. The fairlearn package does not, but the related *fair regressions* method does (Agarwal et al., 2019). This is sufficient for statistical parity (Def. 1), but other parities such as BFPC and BFNC (Def. 2) require further modifications as discussed in § 5

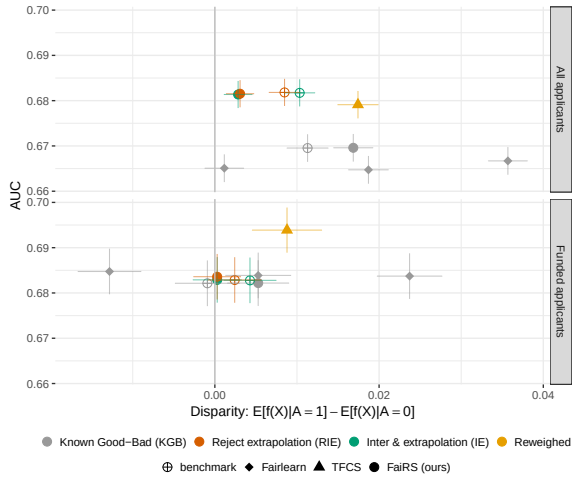

Figure 1. Area under the ROC curve (AUC) with respect to the synthetic outcome against disparity in the average risk prediction for the disadvantaged ($A_i = 1$) vs advantaged ($A_i = 0$) groups. FaiRS reduces disparities for the RIE and IE approaches while maintaining AUCs comparable to the benchmark models (first row). Evaluation on only funded applicants (second row) overestimates the performance of TFCS and KGB models and underestimates disparities for all models. Error bars show the 95% confidence intervals. See § 7 for details.

Communities & Crime dataset, FaiRS improves on statistical parity without compromising performance relative to a benchmark loss-minimizing least squares regression model (See Supplement §D.5).

8. Conclusion

We develop a framework, Fairness in the Rashomon Set (FaiRS), to characterize the range of predictive disparities and find the absolute disparity minimizing model over the set of good models. FaiRS is suitable for a variety of applications including settings with selectively labelled outcomes where the selection decision and outcome are unconfounded given the observed features. The method is generic, applying to both a large class of prediction functions and a large class of predictive disparities.

In many settings, the set of good models is a rich class, in which models differ substantially in terms of their fairness properties. Exploring the range of predictive fairness properties over the set of good models opens new perspectives on how we learn, select, and evaluate machine learning models. A model designer may use FaiRS to select the model with the best fairness properties among the set of good models. FaiRS can be used during evaluation to compare the predictive disparities of a benchmark model against other models in the set of good models. When this evaluation illuminates unjustified disparities in the benchmark model, FaiRS can be used to find a more equitable model with performance comparable to the benchmark. Characterizing the properties of the set of good models is a relevant enterprise for both model designers and auditors alike. This exercise opens new perspectives on algorithmic fairness that may provide exciting opportunities for future research.

Acknowledgements

We are grateful to our data partners at Commonwealth Bank of Australia, and in particular to Nathan Damaj, Bojana Manojlovic and Nikhil Ravichandar for their generous support and assistance throughout this research. We also thank Riccardo Fogliato, Sendhil Mullainathan, Cynthia Rudin and anonymous reviewers for valuable comments and feedback. Rambachan gratefully acknowledges financial support from the NSF Graduate Research Fellowship under Grant No. DGE1745303. Coston gratefully acknowledges financial support support from the K&L Gates Presidential Fellowship and the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE1745016. Any opinions, findings, and conclusions or recommendations expressed in this material are solely those of the authors.

References

Civil rights act, 1964. 42 U.S.C. § 2000e.

Equal credit opportunity act, 1974. 15 U.S.C. § 1691.

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 60–69, 2018.

Agarwal, A., Dudík, M., and Wu, Z. S. Fair regression: Quantitative definitions and reduction-based algorithms. In *Proceedings of the 36th International Conference on Machine Learning, ICML*, 2019.

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., and Rudin, C. Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*, 18: 1–78, 2018.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. there's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016.

Australian Bureau of Statistics. Socio-economic indexes for areas (seifa) technical paper. Technical report, 2016.

Barocas, S. and Selbst, A. Big data's disparate impact. *California Law Review*, 104:671–732, 2016.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Breiman, L. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–215, 2001.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1721–1730, 2015.

Chouldechova, A., Benavides-Prado, D., Fialko, O., and Vaithianathan, R. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. volume 81 of *Proceedings of Machine Learning Research*, pp. 134–148, 2018.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806, 2017.

Coston, A., Ramamurthy, K. N., Wei, D., Varshney, K. R., Speakman, S., Mustahsan, Z., and Chakraborty, S. Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 91–98, 2019.

Coston, A., Mishler, A., Kennedy, E. H., and Chouldechova, A. Counterfactual risk assessments, evaluation and fairness. In *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 582–593, 2020.

Cotter, A., Jiang, H., and Sridharan, K. Two-player games for efficient non-convex constrained optimization. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, pp. 300–332, 2019.

Crook, J. and Banasik, J. Does reject inference really improve the performance of application scoring models? *Journal of Banking & Finance*, 28(4):857–874, 2004.

Dastin, J. Amazon scraps secret ai recruiting tool that showed bias against women, Oct 2018. URL https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Dong, J. and Rudin, C. Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence*, 2(12):810–824, 2020.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. R., and Pontil, M. A. Empirical risk minimization under fairness constraints. In *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 2796–2806, 2018.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Dwork, C., Moritz Hardt, T. P., Reingold, O., and Zemel, R. Fairness through awareness. In *ITCS '12: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268, 2015.

Fisher, A., Rudin, C., and Dominici, F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019. URL http://jmlr.org/papers/v20/18-760.html.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. Predictably unequal? the effects of machine learning on credit markets. Technical report, 2020.

Gillis, T. False dreams of algorithmic fairness: The case of credit pricing. 2019.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 3323–3331, 2016.

Heckman, J. J. Varieties of selection bias. *The American Economic Review*, 80(2):313–318, 1990.

Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., and Wallach, H. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 1–16, 2019.

Imbens, G. W. and Rubin, D. B. *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge, United Kingdom, 2015.

Kallus, N. and Zhou, A. Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning*, pp. 2439–2448, 2018.

Kim, M. P., Ghorbani, A., and Zou, J. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254, 2019.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1): 237–293, 2018.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., and Mullainathan, S. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 275–284, 2017.

Li, Z., Hu, X., Li, K., Zhou, F., and Shen, F. Inferring the outcomes of rejected loans: An application of semisupervised clustering. *Journal of the Royal Statistical Society: Series A*, 183(2):631–654, 2020.

Little, R. J. and Rubin, D. B. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

Mancisidor, R. A., Kampffmeyer, M., Aas, K., and Jenssen, R. Deep generative models for reject inference in credit scoring. *Knowledge-Based Systems*, pp. 105758, 2020.

Marx, C., Calmon, F., and Ustun, B. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 6765–6774, 2020.

Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pp. 107–118. 2018.

Mishler, A. Modeling risk and achieving algorithmic fairness using potential outcomes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 555–556, 2019.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. Technical report, arXiv Working Paper, arXiv:1811.07867, 2019.

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

Nguyen, H.-T. et al. Reject inference in application scorecards: evidence from france. Technical report, University of Paris Nanterre, EconomiX, 2016.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 5680–5689. 2017.

Raghavan, M., Barocas, S., Kleinberg, J., and Levy, K. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 469–481, 2020.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019.

Rudin, C., Wang, C., and Coker, B. The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2(1), 2020.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 59–68, 2019.

Semenova, L., Rudin, C., and Parr, R. A study in rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. Technical report, arXiv preprint arXiv:1908.01755, 2020.

Singh, H., Singh, R., Mhasawade, V., and Chunara, R. Fairness violations and mitigation under covariate shift. In *FAccT'21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.

Vigdor, N. Apple card investigated after gender discrimination complaints, Nov 2019. URL https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 325–333, 2013.

Zeng, G. and Zhao, Q. A rule of thumb for reject inference in credit scoring. *Math. Finance Lett.*, 2014:Article–ID, 2014.