Characterizing Fairness Over the Set of Good Models Under Selective Labels

Amanda Coston[*] **Ashesh Rambachan**[†] **Alexandra Chouldechova**[‡]

arXiv:2101.00352v2 [cs.LG] 13 Jan 2021

January 14, 2021

Abstract

Algorithmic risk assessments are increasingly used to make and inform decisions in a wide variety of high-stakes settings. In practice, there is often a multitude of predictive models that deliver similar overall performance, an empirical phenomenon commonly known as the "Rashomon Effect." While many competing models may perform similarly overall, they may have different properties over various subgroups, and therefore have drastically different predictive fairness properties. In this paper, 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." We provide tractable algorithms to compute the range of attainable group-level predictive disparities and the disparity minimizing model over the set of good models. We extend our framework to address the empirically relevant challenge of selectively labelled data in the setting where the selection decision and outcome are unconfounded given the observed data features. We illustrate our methods in two empirical applications. In a real world credit-scoring task, we build a model with lower predictive disparities than the benchmark model, and demonstrate the benefits of properly accounting for the selective labels problem. In a recidivism risk prediction task, we audit an existing risk score, and find that it generates larger predictive disparities than any model in the set of good models.

1 Introduction

Algorithmic risk assessments are increasingly used to make and inform decisions in a variety of high-stakes settings ranging from health care, the child welfare and criminal justice systems to consumer lending and hiring [1, 2, 3, 4, 5, 6, 7, 8, 9]. A key concern is whether decisions based on these risk assessments can disproportionately harm or advantage sensitive or protected groups. Consequently, there is widespread interest among researchers, practitioners and regulators in both 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. However, for real-world settings where a risk assessment is already in use, practitioners and auditors may instead be interested in assessing disparities with respect to the current model, which we term the *benchmark model*. For example, financial institutions have existing credit score models that are used to decide which applicants should be approved for a loan. When such a benchmark model exists, the relevant

[*]Carnegie Mellon University, Heinz College and Machine Learning Department: acoston@andrew.cmu.edu

[†]Harvard University, Department of Economics: asheshr@g.harvard.edu

[‡]Carnegie Mellon University, Heinz College: achoulde@andrew.cmu.edu

question for the decision-maker is: can we improve upon the benchmark model in terms of predictive fairness with minimal change in overall accuracy?

We provide a procedure for answering this question motivated by the common empirical phenomenon known as the "Rashomon Effect" [10]. The Rashomon Effect describes a setting in which many models perform similarly overall. While many models may perform similarly overall, they may perform quite differently over various subgroups in the population, and therefore have drastically different predictive fairness properties [11]. For models that perform similarly to a chosen benchmark model, which we refer to as *the set of good models* as in [19], we ask the following questions: What are the range of predictive disparities that could be generated over this set? How can we efficiently identify the absolute disparity minimizing model within the set?

These questions are especially relevant for legal audits of disparate impact. In domains such as credit lending and hiring, disparate impact must be justified by "business necessity" [12, 13, 14]. For example, in credit lending, regulators investigate whether the lender could have offered more loans to minority applicants without affecting default rates [15]. In hiring, regulators investigate whether resume screening software screens out underrepresented applicants for reasons that cannot be attributed to the job criteria [9]. Our methods provide one possible technical formalization of the business necessity criteria. An auditor can use our method to assess whether there exists an alternative model within an acceptable model class that reduces predictive disparities without compromising performance. If possible, then it is difficult to justify the benchmark model on the grounds of business necessity.

A key empirical challenge in domains such as hiring and credit lending is that outcomes (i.e., job performance or loan repayment) are not observed for candidates who were not selected—that is, those who were not hired, or whose loans were not ultimately funded [16, 6]. This *selective labels problem* is particularly challenging in the context of assessing predictive fairness. Focusing on only the cases for which outcomes (labels)[4] are available, as is common in analyses of standard benchmark datasets including the UCI adult income and German credit datasets, can provide a misleading characterization of predictive disparities. We provide an extension of our framework that addresses the challenges of selectively labelled data in contexts where the selection decision and outcome are unconfounded given the observed data features.

Contributions: We provide a general method for investigating a rich class of predictive fairness properties over a set of models with comparable overall performance. We (1) demonstrate how this method improves on the fairness properties of a benchmark model without compromising on performance and provide theoretical results to bound the generalization error and predictive disparities of our method [§ 4]; (2) detail the necessary modifications to our procedure for settings with selective labels [§ 5]; (3) use our framework on a real world credit-scoring task to build a model with lower predictive disparities than the benchmark model, and demonstrate the benefits of properly accounting for the selective labels problem [§ 6]; and (4) use our framework to audit the COMPAS risk score for bias, finding that the COMPAS risk score generates larger disparities between black and white defendants than any model in the set of good models [§ 7]. All proofs are given in the Supplement.

2 Background and Related Work

2.1 Rashomon Sets and Predictive Multiplicity

Supervised machine learning algorithms are designed to return the most accurate prediction function given the observable data. However, it is common in practice to observe that many different prediction functions are similarly accurate on average. This empirical phenomenon was famously called the "Rashomon Effect" in [10]. Even though these models may have similar overall accuracy, they may differ along several other dimensions. Recent work on simplicity, interpretability and explainability investigates the properties of models that achieve similar overall accuracy [17, 18, 19, 20, 21].

[4] We use labels and outcomes interchangeably.

We introduce these ideas into research on algorithmic fairness, providing computational techniques to investigate the range of predictive disparities that may be generated over the set of good models. Even though many models deliver similar average accuracy, these models may have drastically different properties over subgroups within the population. We document this phenomenon empirically in leading applications such as credit scoring and recidivism risk prediction.[5] Additionally, we show that the reductions approach developed in [22, 23] may be used to solve optimization problems over the set of good models. This is a general insight that may be of interest to researchers investigating the implications of the Rashomon Effect for other model properties.

2.2 Fair Classification and Fair Regression

Our methods are related to a large, influential literature on fair classification and fair regression [24, 25, 26, 27, 28, 22, 23, 29], which aims to construct prediction functions that minimize average loss subject to a fairness constraint chosen by the decision-maker. In contrast, we construct prediction functions that minimize or maximize a chosen measure of predictive disparities subject to a constraint on overall performance.

While our focus on a benchmark model resembles "post-processing" techniques, there are important differences in both the motivation and methodology. Post-processing techniques typically modify an existing model to achieve a target notion of fairness [26, 30, 31]. However in many settings, decision-makers may find it difficult to exactly specify what level of disparities are acceptable, but may instead know how much performance loss they are willing to tolerate. For example in consumer lending, a financial institution may be unable to specify what types of differences in risk scores across groups are acceptable, but may instead be able to specify an acceptable average default rate among approved loans. Additionally, our methods only use the existing model to calibrate the performance constraint, but need not share any other properties in common with the existing model. Post-processing techniques also require access to the benchmark model's predictions, whereas our method only requires access to the benchmark model's overall performance.

Our work builds upon the reductions approach methodology in [22, 23]. We extend the reductions approach to solve general optimization problems over the set of good models. [29] also provides methods for selecting a classifier that minimizes a particular notion of predictive fairness ("decision boundary covariance") subject to a constraint on its performance. Relative to this work, we allow for any notion of predictive disparities from a large class that encompasses violations of many well-known definitions of predictive fairness such as statistical parity, balance for the positive and negative class and bounded group loss. Furthermore, our results cover both classification and regression tasks. Lastly, we emphasize that our methods achieve a different goal. We directly search for the disparity minimizing model among the set of good models, whereas existing approaches search for the most accurate model among those that satisfy a predictive fairness constraint. As we discuss below, existing methods also do not explicitly handle selectively labeled data in the fair regression or classification framework.

2.3 Selective Labels and Reject Inference

In many risk assessment settings, the training data only contain labeled outcomes for a selectively observed sample of observations from the full population of interest. This *selective labels problem* [16, 6] is common in many settings such as the criminal justice system, hiring and consumer lending. In consumer lending, for instance, financial institutions use risk scores to assess *all* loan applicants, yet the historical data only contains default/repayment outcomes for those applicants whose loans were approved.

A possible solution to the selective labels problem would be to treat the selectively labelled population as if it were the population of interest, and proceed with training and evaluation on the selectively labelled population only. This

[5] Our empirical results on recidivism risk prediction mirror the findings in [11], which show that there exists several models that deliver similar overall performance but have drastically different predictive fairness properties between black and white defendants in the ProPublica COMPAS recidivism data.

solution is termed the "*known good-bad*" (KGB) approach in the credit scoring literature because it restricts attention to only observations with known outcomes [32, 33]. 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 [34, 35].

In contrast, a large literature advocates for various "reject inference" procedures that incorporate information from rejected applications in model construction. Reject inference procedures involve a class of methods for training a risk assessment on imputed outcomes for all applicants using augmentation, reweighing and extrapolation-based approaches [36, 37]. There is considerable debate as to the performance improvements of reject inference over simply training a model on the observed population, but reject inference methods remain popular in credit scoring settings [38, 39, 32, 33]. In our experiments on real-world credit lending data, we found that extrapolation approaches outperformed the KGB approach. Our proposed approach accommodates reject inference methods and, crucially, evaluates fairness over the full population of interest. Finally, a related work is [40], which studies fairness properties in an online classification problem that suffers from the selective labels problem.

3 Setting and Problem Formulation

We consider a setting in which 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 the label. The labels $\mathcal{Y}$ may be either discrete or continuous.

The observed 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 label is observed in the training data if and only if the decision satisfies $D_i \in \mathcal{D}^*$. Hence, the training data are $\{(X_i, A_i, D_i, Y_i 1\{D_i \in \mathcal{D}^*\})\}_{i=1}^n$. For example, in our application to consumer lending decisions, we only observe whether a loan applicant defaulted if her application was approved and funded.

Given a specified set of prediction functions $\mathcal{F}$ with elements $f\colon \mathcal{X} \to [0, 1]$, we wish to find the prediction function $f \in \mathcal{F}$ that

(1) minimizes or maximizes a measure of the predictive disparity between values of the protected or sensitive attribute,

(2) minimizes a measure of the absolute predictive disparity between values of the protected or sensitive attribute

subject to a constraint on the overall performance of the prediction function. We measure overall performance using average loss, where $l\colon \mathcal{Y} \times [0, 1]$ 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 [23].[6] The constraint on overall performance takes the form

$$\text{loss}(f) \leq \epsilon \tag{1}$$

for some specified *loss tolerance* $\epsilon \in [0, 1]$. We refer to the set of prediction functions that satisfy this constraint on average loss as the **set of good models**.

In practice, the decision-maker often has an existing benchmark model $\tilde{f}$. In criminal justice and consumer lending settings, there may be an existing risk score in use. In this case, the decision-maker may define $\epsilon = (1 + \delta)\,\text{loss}(\tilde{f})$ for $\delta \in [0, 1]$. The set of good models is then the set of prediction functions $f \in \mathcal{F}$ whose average performance lies within a δ-neighborhood of the benchmark model $\tilde{f}$. When the loss tolerance is calibrated in this manner, the set of good models is referred to as the "Rashomon set" in [17, 18, 19, 20].

[6] That is, for all $y, \tilde{y} \in \mathcal{Y}$ and $u, \tilde{u} \in [0, 1]$, $|l(y, u) - l(\tilde{y}, \tilde{u})| \leq |y - \tilde{y}| + |u - \tilde{u}|$. As noted in [23], this assumption is satisfied by the least-squares loss $l(y, u) = (y - u)^2/2$ for $\mathcal{Y} \in [0, 1]$ and the logistic regression loss $l(y, u) = \log(1 + e^{-C(2y-1)(2u-1)})/(2\log(1 + e^C))$ for $C > 1$ and $\mathcal{Y} = \{0, 1\}$.

3.1 Measures of Predictive Disparities

We consider measures of predictive disparity that can be written in the form

$$\text{disparity}(f) := \beta_0 \mathbb{E}\left[f(X_i) | \mathcal{E}_0\right] + \beta_1 \mathbb{E}\left[f(X_i) | \mathcal{E}_1\right], \tag{2}$$

where $\mathcal{E}_a$ 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 (2) defines the predictive disparity over the *full* population of interest (i.e., not conditional on the decision).

For different choices of the conditioning events $\mathcal{E}_0, \mathcal{E}_1$ and parameters β_0, β_1, (2) encodes violations of commonly used definitions of predictive fairness.

Example 1 (Statistical Parity)**.** ***Statistical parity*** *requires the prediction $f(X_i)$ to be statistically independent of the attribute A_i [24, 25, 41]. By setting $\mathcal{E}_a = \{A_i = a\}$ for $a \in \{0,1\}$ and $\beta_0 = -1, \beta_1 = 1$, (2) measures a particular violation of statistical parity, where*

$$\text{disparity}(f) = \mathbb{E}\left[f(X_i) \mid A_i = 1\right] - \mathbb{E}\left[f(X_i) \mid A_i = 0\right]$$

is the average difference in predictions across values of the attribute. Notice that $\text{disparity}(f) = 0$ implies the prediction is mean-independent of the attribute (i.e., $\mathbb{E}\left[f(X_i) \mid A_i = 0\right] = \mathbb{E}\left[f(X_i) \mid A_i = 1\right]$).

Example 2 (Balance for the Positive Class and Negative Class)**.** *Suppose $\mathcal{Y} = \{0,1\}$.* ***Balance for the positive class*** *and* ***balance for the negative class*** *requires the prediction $f(X_i)$ to be statistically independent of the attribute A_i conditional on $Y_i = 1$ and $Y_i = 0$ respectively (e.g., see Chapter 2 of [42]). Define $\mathcal{E}_a = \{Y_i = 1, A_i = a\}$ for $a \in \{0,1\}$ and $\beta_0 = -1, \beta_1 = 1$. Then, (2) measures a particular violation of balance for the positive class*

$$\text{disparity}(f) = \mathbb{E}\left[f(X_i) \mid Y_i = 1, A_i = 1\right] - \mathbb{E}\left[f(X_i) \mid Y_i = 1, A_i = 0\right]$$

by summarizing the average difference in predictions across values of the attribute given $Y_i = 1$. If instead we define $\mathcal{E}_a = \{Y_i = 0, A_i = a\}$ for $a \in \{0,1\}$, then (2) equals

$$\text{disparity}(f) = \mathbb{E}\left[f(X_i) \mid Y_i = 0, A_i = 1\right] - \mathbb{E}\left[f(X_i) \mid Y_i = 0, A_i = 0\right],$$

which is the average difference in predictions across values of the attribute given $Y_i = 0$ and measures an average violation of balance for the negative class.

While our focus on differences in average predictions across values of the attribute is weaker than requiring full statistical independence, it will enable us to develop tractable computational procedures. Additionally, focusing on average differences in predictions across groups is a common relaxation of such parity-based predictive fairness definitions [43, 44].

The definition of predictive disparity (2) also accommodates *fairness promoting interventions*, which aim to increase opportunities for a particular group. In a particular application, it may be reasonable to assume that the decision-maker is more likely to assign the desired decision to cases with larger predicted values $f(X_i)$. Hence, the decision-maker may wish to search for the prediction function among the set of good models that maximizes the average predicted value $f(X_i)$ for a particular group.

Example 3 (Fairness Promoting Interventions)**.** *Defining $\mathcal{E}_1 = \{A_i = 1\}$ and $\beta_0 = 0, \beta_1 = 1$,*

$$\text{disparity}(f) = \mathbb{E}\left[f(X_i) \mid A_i = 1\right]$$

measures the average risk score for the group with $A_i = 1$. We refer to this choice as an ***affirmative action****-based fairness promoting intervention.*

Further assuming $\mathcal{Y} = \{0, 1\}$ and defining $\mathcal{E}_1 = \{Y_i = 1, A_i = 1\}$,

$$\text{disparity}(f) = \mathbb{E}\left[f(X) \mid Y_i = 1, A_i = 1\right]$$

measures the average risk score for the group with both $Y_i = 1$, $A_i = 1$. We refer to this choice as a ***qualified affirmative action****-based fairness promoting intervention.*

In the Supplement, we show that our analysis extends to *bounded group loss*, which is another notion of predictive fairness that requires that the average loss conditional on each value of the attribute achieve some threshold [23].

3.2 Characterizing Predictive Disparities over the Set of Good Models

First, we wish to characterize the range of predictive disparities over the set of good models. To do so, we search for the prediction function $f \in \mathcal{F}$ that minimizes or maximizes the predictive disparity measure subject to a constraint on average loss

$$\min_{f \in \mathcal{F}} \text{disparity}(f) \text{ subject to } \text{loss}(f) \leq \epsilon, \tag{3}$$

$$\max_{f \in \mathcal{F}} \text{disparity}(f) \text{ subject to } \text{loss}(f) \leq \epsilon. \tag{4}$$

Since the maximization problem can be recast as a minimization problem, we focus on the minimization problem without loss of generality. Second, we wish to find the prediction function that minimizes the absolute predictive disparity over the set of good models

$$\min_{f \in \mathcal{F}} |\text{disparity}(f)| \text{ subject to } \text{loss}(f) \leq \epsilon. \tag{5}$$

These exercises are useful to decision-makers that wish to construct fair prediction functions and auditors that wish to evaluate existing prediction functions. For auditors, (3) enables them to trace out the range of predictive disparities that *could* be generated in a given setting, and thereby identify where the existing prediction function lies on this frontier. This relates 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 [12, 13, 14]. For decision-makers, (5) enables them to search for prediction functions that reduce absolute predictive disparities without compromising predictive performance.

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

The problems of characterizing the range of predictive disparities (3) and finding the absolute predictive disparity minimizing model (5) over the set of good models may be tractably solved through a reductions approach, as developed in [22, 23].

In this section, we focus on the simplest case in which there is no selective labels problem, meaning $\mathcal{D}^* = \mathcal{D}$ and the label Y_i is observed for all observations in the training data. In § 5, we extend our analysis to the empirically relevant selective labels case with $\mathcal{D}^* \subset \mathcal{D}$.

In the main text, we present the reductions approach in detail to solve (3). In the Supplement, we extend the reductions approach to solve for the absolute disparity minimizing model (5), obtaining analogous performance guarantees.

4.1 Computing the Range of Predictive Disparities

Following the reductions approach in [22, 23], we consider randomized prediction functions that select $f \in \mathcal{F}$ according to some distribution $Q \in \Delta(\mathcal{F})$. Let $\text{loss}(Q) := \sum_{f \in \mathcal{F}} Q(f) \, \text{loss}(f)$ and $\text{disparity}(Q) := \sum_{f \in \mathcal{F}} Q(f) \, \text{disparity}(f)$.

Our goal is to construct solutions to

$$\min_{Q\in\Delta(\mathcal{F})} \text{disparity}(Q) \text{ subject to } \text{loss}(Q) \leq \epsilon. \tag{6}$$

In the reductions approach, a key object 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 proceed by first reducing the optimization problem of interest to a constrained classification problem through a discretization argument, and second solving the resulting constrained classification problem through a further reduction to finding the equilibrium in a min-max game.

4.1.1 Reduction to Constrained Classification

We construct discrete approximations to the loss function and disparity measure following the argument and notation in [23].

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 $\mathcal{Y}$.[7] Define a piecewise approximation to the loss function as $l_\alpha(y,u) := l(\underline{y}, [u]_\alpha + \frac{\alpha}{2})$, 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. Since the loss function is 1-Lipshitz, it follows that $|l(y,u) - l_\alpha(y,u)| \leq \alpha$, and $\text{loss}(f) \leq \text{loss}_\alpha(f) + \alpha$ for any $f \in \mathcal{F}$, where $\text{loss}_\alpha(f) := \mathbb{E}\left[l_\alpha(Y_i, f(X_i))\right]$. Therefore, for a fine enough discretization grid, $\text{loss}_\alpha(f)$ provides a high-quality approximation to $\text{loss}(f)$.

Next, define

$$c(y,z) := N \times \Big(l(y, z + \frac{\alpha}{2}) - l(y, z - \frac{\alpha}{2})\Big) \tag{7}$$

and Z_α to be the random variable that samples $z_\alpha \in \mathcal{Z}_\alpha$ uniformly at random and is independent of the data (X_i, A_i, Y_i). For any $h_f \in \mathcal{H}$, define the cost-sensitive average loss function as

$$\text{cost}(h_f) := \mathbb{E}\left[C(\underline{Y}_i, Z_\alpha) h_f(X_i, Z_\alpha)\right]. \tag{8}$$

We apply Lemma 1 in [23] to relate $\text{cost}(h_f)$ to $\text{loss}_\alpha(f)$.

Lemma 1. *Given any distribution over (X_i, A_i, Y_i) and $f \in \mathcal{F}$,*

$$\text{cost}(h_f) + c_0 = \text{loss}_\alpha(f),$$

where $c_0 \geq 0$ is a constant that does not depend on f.

Since $\text{loss}_\alpha(f)$ provides a high-quality approximation for $\text{loss}(f)$, Lemma 1 implies that $\text{cost}(h_f)$ also provides a high-quality approximation for $\text{loss}(f)$. For any $Q \in \Delta(\mathcal{F})$, define $Q_h \in \Delta(\mathcal{H})$ to be the induced distribution over threshold classifiers h_f. Furthermore, Lemma 1 implies $\text{cost}(Q_h) + c_0 = \text{loss}_\alpha(Q)$, 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.

Next, we relate the predictive disparity measure defined on prediction functions $f \in \mathcal{F}$ to a predictive disparity measure defined on threshold classifiers $h_f \in \mathcal{H}$. Towards this, define

$$\text{disparity}(h_f) := \beta_0 \mathbb{E}\left[h_f(X_i, Z_\alpha) \mid \mathcal{E}_0\right] + \beta_1 \mathbb{E}\left[h_f(X_i, Z_\alpha) \mid \mathcal{E}_1\right]. \tag{9}$$

Lemma 2. *Given any distribution over (X_i, A_i, Y_i) and $f \in \mathcal{F}$,*

$$|\text{disparity}(h_f) - \text{disparity}(f)| \leq (|\beta_0| + |\beta_1|)\,\alpha.$$

[7] That is, $\tilde{\mathcal{Y}}_\alpha \subseteq \mathcal{Y}$ and satisfies (1) for any $y \in \mathcal{Y}$, there exists $\tilde{y} \in \tilde{\mathcal{Y}}_\alpha$ with $|y - \tilde{y}| \leq \frac{\alpha}{2}$, and (2) for any $\tilde{y}, \tilde{y}' \in \tilde{\mathcal{Y}}_\alpha$, $|\tilde{y} - \tilde{y}'| > \frac{\alpha}{2}$. If $\mathcal{Y} = \{0,1\}$, then we may simply define $\tilde{\mathcal{Y}}_\alpha = \{0,1\}$.

Hence, the predictive disparity measure defined over threshold classifiers $h_f \in \mathcal{H}$ approximates the predictive disparity measure defined over prediction functions $f \in \mathcal{F}$. Lemma 2 and Jensen's Inequality imply $|\operatorname{disparity}(Q_h) - \operatorname{disparity}(Q)| \leq (|\beta_0| + |\beta_1|)\,\alpha$.

Lemmas 1-2 suggest that we may approximate (6) with a discretized problem over threshold classifiers

$$\min_{Q_h \in \Delta(\mathcal{H})} \operatorname{disparity}(Q_h) \text{ subject to } \operatorname{cost}(Q_h) \leq \epsilon - c_0. \tag{10}$$

We construct solutions to the empirical analogue

$$\min_{Q_h \in \Delta(\mathcal{H})} \widehat{\operatorname{disparity}}(Q_h) \text{ subject to } \widehat{\operatorname{cost}}(Q_h) \leq \hat{\epsilon}, \tag{11}$$

where $\hat{\epsilon} := \epsilon - \hat{c}_0$ plus additional slack, and $\hat{c}_0, \widehat{\operatorname{disparity}}(Q_h), \widehat{\operatorname{cost}}(Q_h)$ are the associated sample analogues.

4.1.2 The Algorithm

We solve (11) by forming its Lagrangian with primal variable $Q_h \in \Delta(\mathcal{H})$ and dual variable $\lambda \in \mathbb{R}^+$

$$L(Q_h, \lambda) := \widehat{\operatorname{disparity}}(Q_h) + \lambda(\widehat{\operatorname{cost}}(Q_h) - \hat{\epsilon}). \tag{12}$$

Solving (11) is equivalent to finding the saddle point of the min-max problem

$$\min_{Q_h \in \Delta(\mathcal{H})} \max_{0 \leq \lambda \leq B_\lambda} L(Q_h, \lambda), \tag{13}$$

where $B_\lambda \geq 0$ bounds the Lagrange multiplier. We search for the saddle point by treating it as the equilibrium of a zero-sum game between the Q_h-player and λ-player, extending the exponentiated gradient algorithm proposed in [22, 23]. The algorithm delivers a ν-approximate saddle point of the Lagrangian, which is a pair $(\hat{Q}_h, \hat{\lambda})$ satisfying

$$L(\hat{Q}_h, \hat{\lambda}) \leq L(Q_h, \hat{\lambda}) + \nu \text{ for all } Q_h \in \Delta(\mathcal{H}), \tag{14}$$

$$L(\hat{Q}_h, \hat{\lambda}) \geq L(\hat{Q}_h, \lambda) - \nu \text{ for all } 0 \leq \lambda \leq B_\lambda. \tag{15}$$

Algorithm 1 implements the exponentiated gradient algorithm, except for the best-response functions of the λ-player and the Q_h-player. The best-response function of the λ-player is

$$\operatorname{Best}_\lambda(Q_h) := \begin{cases} 0 \text{ if } \widehat{\operatorname{cost}}(Q_h) - \hat{\epsilon} \leq 0, \\ B_\lambda \text{ otherwise.} \end{cases} \tag{16}$$

The best-response function of the Q_h-player may be constructed through a further reduction to cost-sensitive classification. To see this, notice that the Lagrangian may be re-written as

$$L(h_f, \lambda) = \hat{\mathbb{E}}\left[\mathbb{E}_{Z_\alpha}\left[c_\lambda(\underline{Y}_i, A_i, Z_\alpha) h_f(X_i, Z_\alpha)\right]\right] - \lambda\hat{\epsilon}, \tag{17}$$

where

$$c_\lambda(\underline{Y}_i, A_i, Z_\alpha) := \tfrac{\beta_0}{\hat{p}_0} 1\{\mathcal{E}_0\} + \tfrac{\beta_1}{\hat{p}_1} 1\{\mathcal{E}_1\} + \lambda c(\underline{Y}_i, Z_\alpha) \tag{18}$$

$$\hat{p}_a := \hat{\mathbb{E}}\left[\mathcal{E}_a\right] \tag{19}$$

for $a \in \{0, 1\}$. We solve this by calling a cost-sensitive classification oracle on an augmented dataset of size $n \times N$ with observations $\{(X_{i,z_\alpha}, C_{i,z_\alpha}\}_{i \in [n], z_\alpha \in \mathcal{Z}_\alpha}$ with $X_{i,z_\alpha} = (X_i, z_\alpha)$ and $C_{i,z_\alpha} = c_\lambda(\underline{Y}_i, A_i, z_\alpha)$.[8]

[8] In our empirical implementation, we use the heuristic least-squares reduction described in [23], which eases the computational burden of the algorithm. We found that the heuristic reduction generally performed well in our empirical work, but performance losses depended on the dataset and the choice of predictive disparity.

Algorithm 1: Algorithm for finding the predictive disparity minimizing model over the set of good models

Input: Training data $\{(X_i, Y_i, A_i)\}_{i=1}^n$,
Parameters β_0, β_1, Events $\mathcal{E}_0, \mathcal{E}_1$, and empirical loss tolerance $\hat{\epsilon}$
Bound B_λ, accuracy ν and learning rate η
Result: ν-approximate saddle point $(\hat{Q}_h, \hat{\lambda})$
Set $\theta_1 = 0 \in \mathbb{R}$;
for $t = 1, 2, \ldots$ **do**
  Set $\lambda_t = B_\lambda \frac{\exp(\theta_t)}{1+\exp(\theta_t)}$;
  $h_t \leftarrow \text{Best}_h(\lambda_t)$;
  $\hat{Q}_{h,t} \leftarrow \frac{1}{t}\sum_{s=1}^t h_s, \quad \bar{L} \leftarrow L(\hat{Q}_{h,t}, \text{Best}_\lambda(\hat{Q}_{h,t}))$;
  $\hat{\lambda}_t \leftarrow \frac{1}{t}\sum_{s=1}^t \lambda_s, \quad \underline{L} \leftarrow L(\text{Best}_h(\hat{\lambda}_t), \hat{\lambda}_t)$;
  $\nu_t \leftarrow \max\left\{L(\hat{Q}_{h,t}, \hat{\lambda}_t) - \underline{L}, \bar{L} - L(\hat{Q}_{h,t}, \hat{\lambda}_t)\right\}$;
  if $\nu_t \leq \nu$ **then**
    if $\widehat{\text{cost}}(\hat{Q}_{h,t}) \leq \hat{\epsilon} + \frac{|\beta_0|+|\beta_1|+2\nu}{B_\lambda}$ **then**
      return $(\hat{Q}_{h,t}, \hat{\lambda}_t)$;
    else
      return *null*
    end
  end
  Set $\theta_{t+1} = \theta_t + \eta\left(\widehat{\text{cost}}(h_t) - \hat{\epsilon}\right)$;
end

4.1.3 Error Analysis

We analyze the suboptimality of the solution returned by Algorithm 1 in the original problem (3), which can be controlled under assumptions about the complexity of the model class $\mathcal{F}$ and conditions on how the parameters in Algorithm 1 are set.

Assumption 1. *Let $R_n(\mathcal{H})$ be the Radermacher complexity of $\mathcal{H}$. Assume 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 $C'' \geq \sqrt{\frac{-\log(\delta/8)}{2}}$.*

Then, Algorithm 1 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. *$\hat{Q}_h \neq null$ and for any $\tilde{Q}$ that is feasible in (6)*

$$\text{loss}(\hat{Q}_h) \leq \epsilon + \tilde{O}(n^{-\phi}),$$
$$\text{disparity}(\hat{Q}_h) \leq \text{disparity}(\tilde{Q}) + \tilde{O}(n_0^{-\phi}) + \tilde{O}(n_1^{-\phi}),$$

   where n_0, n_1 are the number of samples satisfying the events $\mathcal{E}_0, \mathcal{E}_1$ respectively and the notation $\tilde{O}(\cdot)$ suppresses polynomial dependence on $\ln(n)$ and $\ln(1/\delta)$.

2. *$\hat{Q}_h = null$ and (6) is infeasible.*

Infeasibility is only a concern in settings in which there does not exist a prediction function $f \in \mathcal{F}$ that satisfies the average loss constraint. This result extends Theorem 3 in [22] and Theorem 2 in [23], showing that the solution $\hat{Q}_h$ returned by Algorithm 1 is approximately feasible and achieves the lowest possible predictive disparity up to some error.

4.2 Shrinking the Support of the Stochastic Risk Score

A key challenge to the practical use of Algorithms 1 is it returns a stochastic prediction function $\hat{Q}_h$ with possibly large support. The number of prediction functions in the support of $\hat{Q}_h$ is equal to the total number of iterations taken by the respective algorithm. As a result, $\hat{Q}_h$ may be complex to describe, time-intensive to evaluate, and memory-intensive to store.

The support of the returned stochastic prediction may be shrunk while maintaining the same guarantees on its performance by solving a simple linear program. To do so, we take the set of prediction functions in the support of $\hat{Q}_h$ and solve the following linear program

$$\min_{p \in \Delta^T} \sum_{t=1}^{T} p_t \widehat{\text{disparity}}(h_t) \text{ subject to } \sum_{t=1}^{T} p_t \widehat{\text{cost}}(h_t) \leq \hat{\epsilon} + 2\nu, \tag{20}$$

where T is the number of iterations of Algorithm 1, Δ^T is the T-dimensional unit simplex and h_t is the t-th prediction function in the support of $\hat{Q}_h$ (i.e., the prediction function constructed at the t-th iteration of Algorithm 1). We then use the randomized prediction function that assigns probability p_t to each prediction function in the support of $\hat{Q}_h$.[9]

Lemma 7 of [45] shows that the solution to (20) has at most 2 support points and the same performance guarantees as the original solution $\hat{Q}_h$. We use the linear programming reduction when evaluating performance on held out test data in our empirical experiments.

5 Optimizing Over the Set of Good Models Under Selective Labels

In this section, we extend the reductions approach to the case in which the training data suffer from the selective labels problem, meaning that the outcome Y_i is observed only if $D_i \in \mathcal{D}^*$ with $\mathcal{D}^* \subset \mathcal{D}$. The selective labels problem arises in many empirical applications such as consumer lending decisions. The error analysis of Algorithm 1 continues to hold under selective labels provided there is oracle access to the true regression function.

5.1 The Selective Labels Problem

Consider a binary decision setting with $\mathcal{D} = \{0, 1\}$ and $\mathcal{D}^* = \{1\}$. The outcome Y_i is only observed if $D_i = 1$ and is otherwise unobserved. Without further assumptions, average loss and measures of predictive disparity (2) that condition on the outcome Y_i are not identified due to the selective labels problem. This poses a significant challenge as we are interested in characterizing and evaluating model properties in (3) and (5) over the full population.

To make progress, we introduce the following assumption on the nature of the selective labels problem.[10]

Assumption 2. *Assume the population distribution $(X_i, A_i, D_i, Y_i) \sim P$ satisfies*

1. ***Selection on observables**: $D_i \perp\!\!\!\perp Y_i \mid X_i$.*
2. ***Positivity**: $\mathbb{P}\left(D_i = 1 \mid X_i = x\right) > 1$ with probability one.*

Assumption 2 is common in causal inference and selection problem settings (e.g., see Chapter 12 of [46] and [47]). Selection on observables requires that the observed features X_i measure captures all features that jointly influence the decision D_i and the outcome Y_i. Positivity requires that each observation has a positive probability of receiving the decision $D_i = 1$.

[9] In practice, we calibrate the constraint in (20) by choosing the smallest $\nu \geq 0$ such that the linear program has a feasible solution, following the practical recommendations in [45].

[10] It may be useful to cast this setting into a potential outcomes framework, where Y_i^D denotes the outcome that would be observed if decision $D = 1$ were assigned. Since we are primarily in a pure selection problem, we may define $Y_i^{D=0} = 0$ (e.g., a loan application that is not approved and funded cannot default by definition). The observed outcome Y_i then equals $Y_i^{D=1} D_i$.

Let $\mu(x) := \mathbb{E}[Y_i \mid X_i = x]$ denote the true regression function of interest. Under Assumption 2, $\mu(x)$ is identified since $\mathbb{E}[Y_i \mid X_i] = \mathbb{E}[Y_i \mid X_i, D_i = 1]$. Hence, we may estimate $\mu(x)$ by simply regressing the outcome Y_i on the features X_i among observations with $D_i = 1$. This estimator $\hat{\mu}(x)$ is commonly referred to as a *known-good-bad* (KGB) model in the reject inference literature [32, 33]. In some selective labels settings it may be sufficient to deploy the KGB model; however, when fairness is a concern, restricting attention to the known applicants can be misleading.

5.2 Predictive Disparities under Selective Labels

To see the challenge posed by the selective labels problem, suppose the fairness target is statistical parity. The predictive disparity measure among observations with known outcomes is

$$\mathbb{E}[f(X_i) \mid D_i = 1, A_i = 1] - \mathbb{E}[f(X_i) \mid D_i = 1, A_i = 0]. \tag{21}$$

In contrast, the predictive disparity measure over the full population is $\mathbb{E}[f(X_i) \mid A_i = 1] - \mathbb{E}[f(X_i) \mid A_i = 0]$, which equals

$$\begin{aligned}
&= \mathbb{E}[f(X_i) \mid D_i = 1, A_i = 1]\mathbb{P}(D_i = 1 \mid A_i = 1) \\
&\quad - \mathbb{E}[f(X_i) \mid D_i = 1, A_i = 0]\mathbb{P}(D_i = 1 \mid A_i = 0) \\
&\quad + \mathbb{E}[f(X_i) \mid D_i = 0, A_i = 1]\mathbb{P}(D_i = 0 \mid A_i = 1) \\
&\quad - \mathbb{E}[f(X_i) \mid D_i = 0, A_i = 0]\mathbb{P}(D_i = 0 \mid A_i = 1)
\end{aligned}$$

Depending on the values of the black terms, the population predictive disparity can be quite different from the predictive disparity among only observations with known outcomes [34, 35]. In most settings, we are concerned about disparities in how D is assigned and therefore we care about achieving parity on the *full* population.

It is straightforward to estimate predictive disparity measures (2) for conditioning events $\mathcal{E}$ that only depend on A on the full population. In order to estimate predictive disparity measures for conditioning events $\mathcal{E}$ that additionally depend on Y on the full population, we need Assumption 2 as well as a more general definition of predictive disparity than previously given in § 3 and § 4. To see this, let us now define the modified predictive disparity measure over threshold classifiers as

$$\text{disparity}(h_f) := \beta_0 \frac{\mathbb{E}\left[g(X_i, Y_i) h_f(X_i, Z_\alpha) \mid \mathcal{E}_0\right]}{\mathbb{E}[g(X_i, Y_i) \mid \mathcal{E}_0]} + \beta_1 \frac{\mathbb{E}\left[g(X_i, Y_i) h_f(X_i, Z_\alpha) | \mathcal{E}_1\right]}{\mathbb{E}[g(X_i, Y_i) \mid \mathcal{E}_1]}. \tag{22}$$

For illustrative purposes, we focus on specifying the qualified affirmative action fairness-promoting intervention (Example 3) in this form. Under Assumption 2, we may rewrite the qualified affirmative action disparity measure as[11]

$$\mathbb{E}[f(X_i) | Y_i = 1, A_i = 1] = \frac{\mathbb{E}[f(X_i)\mu(X_i) | A_i = 1]}{\mathbb{E}[\mu(X_i) | A_i = 1]}. \tag{23}$$

This may then be estimated on the full training population by plugging in our estimate of $\hat{\mu}(x)$. With this result, we specify our disparity target over threshold classifiers (22) with $\beta_0 = 0$, $\beta_1 = 1$, $\mathcal{E}_1 = 1\{A_i = 1\}$, and $g(X_i, Y_i) = \hat{\mu}(X_i)$.[12]

5.3 Imputing Missing Outcomes

To estimate loss on the full population, we turn to extrapolation methods in reject inference, which impute outcomes for the unknown cases and then proceed with standard supervised learning.

[11] Derivation in proof of Lemma 9

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

More concretely, *reject inference by extrapolation* (RIE) uses predictions from the KGB model as pseudo-outcomes for the unknown observations [39].[13] We also consider an approach that uses the $\hat{\mu}(x)$ KGB predictions as pseudo-outcomes for *all* applicants, which we call *interpolation & extrapolation* (IE). This method replaces the labelled $\{0, 1\}$ outcomes 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 2-3 summarize the RIE and IE methods for reject inference. If the KGB model could perfectly estimate $\mu(x)$, then the IE approach recovers an oracle setting for which our error analysis for Algorithm 1 continues to hold. Before stating this result formally (§ 5.5), we summarize the modifications necessary to apply our approach in the selective labels setting.

Algorithm 2: Reject inference by extrapolation (RIE) for the selective labels setting

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

Algorithm 3: Interpolation and extrapolation (IE) method for the selective labels setting

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

5.4 Modifications to the Reductions Approach to Accommodate Selective Labels

To account for the selective labels problem in the reductions approach, we must estimate performance and disparity on both the known and unknown cases in our training data. To estimate performance, we use pseudo-outcomes for the unknown cases which can be estimated using either the RIE approach (Algorithm 2) or IE approach (Algorithm 3). To compute disparities of events $\mathcal{E}$ that only depends on A, such as statistical parity or the affirmative action fairness-promoting intervention, we require no further modifications. To compute disparities of events $\mathcal{E}$ that additionally depends on Y such as balance for the positive (or negative) class, we must modify the costs described in (18) as

$$c_\lambda(\underline{\hat{\mu}}_i, A_i, Z_\alpha) = \frac{\beta_0}{\hat{p}} g(X_i, Y_i) 1\{\mathcal{E}_0\} + \frac{\beta_1}{\hat{p}} g(X_i, Y_i) 1\{\mathcal{E}_1\} + \lambda c(\underline{\hat{\mu}}_i, Z_\alpha),$$

where $\hat{p} = \hat{\mathbb{E}}[g(X_i, Y_i)]$. With these modifications, Algorithm 1 may be applied as described in § 4, passing the training data augmented with the imputed outcomes as input.

5.5 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 Algorithm 1 continues to hold under selective labels provided there is oracle access to the true regression function.

[13] One might also consider an inverse-probability weighing (IPW) approach that relies on an estimate $\mathbb{P}(D = 1 \mid X)$ in lieu of an estimate of $\hat{\mu}$. While common in causal inference, this approach is not as common in real-world credit lending settings.

Theorem 2 (Selective Labels)**.** *Suppose Assumption 2 holds and Algorithm 1 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, Algorithm 1 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. $\hat{Q}_h \neq null$ *and for any* $\tilde{Q}$ *that is feasible in (6)*

$$\text{loss}_\mu(\hat{Q}_h) \leq \epsilon + \tilde{O}(n^{-\phi}),$$
$$\text{disparity}(\hat{Q}_h) \leq \text{disparity}(\tilde{Q}) + \tilde{O}(n_0^{-\phi}) + \tilde{O}(n_1^{-\phi}).$$

2. $\hat{Q}_h = null$ *and (6) is infeasible.*

In practice, the estimation error in $\hat{\mu}(x)$ will affect the bounds presented in Theorems 2. The empirical analysis in the next section finds that our method performs well even when using an estimate of $\mu(x)$.

6 Application: Consumer Lending Risk Scores

In this section, we apply our methods to find the absolute predictive disparity minimizing model over the set of good models in a real-world consumer lending context with selectively labeled outcomes. The goal of the following experiment is to demonstrate how a decision maker, when given a set of candidate models, might select the fairest models out of the set of good models We empirically illustrate that it is crucial to account for the selective labels problem in the construction and evaluation of predictive models. Failure to account for selective labels significantly underestimates disparities.

6.1 Data Description

We use data on a sample of personal loan applications to Commonwealth Bank of Australia, a large financial institution in Australia (henceforth, "CommBank"). The sample consists of personal loan applications that were submitted to CommBank from July 2017 to July 2019 by new-to-bank customers (i.e., 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. For example, personal loans are commonly used to purchase a used car and refinance existing debt. In our sample, the median personal loan size AU$ 10,000 and the median interest rate is 13.9% per annum.

For each loan application, we observe all application-level information such as the applicant's reported income and credit score, whether the application was approved by CommBank, the offered terms of the loan, whether the applicant accepted the offered terms and whether the applicant defaulted on the loan. An application is "funded" if it is both approved by CommBank and the offered terms were accepted by the applicant. 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. In our sample of personal loans, 55.5% applications were approved by CommBank and 44.9% of loans were funded. Of the applications that were approved and accepted, only 2.0% of loans defaulted within 5 months.

We focus on predictive disparities across SA4 geographic regions within Australia. SA4 regions are geographic regions defined by the Australian Bureau of Statistics ("ABS") and are roughly analogous to counties in the United States. We define an SA4 region to be socioeconomically disadvantaged based on the ABS' Index of Relative Socioeconomic Disadvantage (IRSD). The IRSD is an index that aggregates census data related to socioeconomic disadvantage (e.g., the fraction of households making less than AU$26,000, the fraction of households with no internet access, etc) into a single index that summarizes relative socioeconomic disadvantage across geographic regions in Australia. Complete details on how the construction of the IRSD may be found in [48]. We say an SA4 region is socioeconomically disadvantaged

($A_i = 1$) if it falls in the top quartile of SA4 regions based on the IRSD.[14] Otherwise, we say an SA4 region is not socioeconomically disadvantaged ($A_i = 0$). The disadvantaged group is relatively under-represented among funded applications as compared to all submitted applications, comprising 21.7% of all loan applications, but only 19.7% of all funded loan applications in our sample.

6.2 Semi-Synthetic Simulation Design

The task is to predict the likelihood of default $Y_i = 1$ based on the information in the loan application X_i. As mentioned, our goal is to illustrate how a decision-maker may find a model with comparable performance to the benchmark model that reduces disparities. We consider three benchmark models: KGB, RIE, and IE. Motivated by a desire to reduce disparities in credit access across geographic regions, we present results for the statistical parity disparity measure between applicants from socioeconomically disadvantaged and non-socioeconomically disadvantaged SA4 regions.

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 using the observed features in the data. On a 20% sample of all applicants, we estimate the functions $\hat{\pi}(x) := \hat{P}(D_i = 1|X_i = x)$ and $\hat{\mu}(x) := \hat{P}(Y_i = 1|X_i = x)$. We then generate synthetic funding decisions $\tilde{D}_i$ and synthetic default outcomes $\tilde{Y}_i$ according to $\tilde{D}_i \mid X_i \sim Bernoulli(\hat{\pi}(X_i))$ and $\tilde{Y}_i \mid X_i \sim Bernoulli(\hat{\mu}(X_i))$. We train all models as if we only knew the synthetic outcome for the synthetically funded applications. The benchmark models are the loss-minimizing KGB, RIE, and IE models, whose average losses over the training data are used to select the corresponding loss tolerances ϵ. Results are reported on all applicants in a held out test set, and performance metrics are constructed with respect to the synthetic outcome $\tilde{Y}$.

6.3 Experimental Results

Figure 1 shows the AUC (y-axis) against disparity (x-axis) for the KGB, RIE, IE benchmarks and disparity-minimizing variants (where each disparity-minimizer corresponds to a different loss tolerance ϵ). The first row shows the evaluation on all applicants in the test set, and the second row shows the metrics for only the funded (known) applicants in the test set. On the target full population, our method is able to reduce disparities for the RIE and IE approaches with negligible loss in performance (first row). The benchmark KGB model is able to achieve such a low disparity on the *funded* applicants that due to generalization error, the "disparity-minimizers" on the train set appear to exacerbate disparities on the test set. (See Appendix E.1 for further details). However, KGB does not account for selective labels, and notably, evaluation on the full population shows that the the KGB disparity is significantly higher than evaluation on known applicants would suggest. Evaluation on the funded cases generally underestimates disparities across the methods. In real-world settings where it is not possible to know the outcomes for all applicants, we recommend a counterfactual evaluation that estimates the counterfactual outcomes for unknown applicants such as that described in [34].

We next consider the extent to which our method can help achieve the goal of increasing credit access for those in the disadvantaged group. Figure 2 shows the distribution of risk scores for the disadvantaged group for the KGB, RIE, and IE models for the benchmark models (first row), and the disparity-minimizer in the set of models with 1% loss tolerance (second row) and 10% loss tolerance (third row). Moving down the left column, which shows the risk distribution on *all* applicants from the disadvantaged group, the risk distribution moves right, indicating that our algorithm reduces the predicted risks for all methods, thereby expanding access to credit for the disadvantaged group.

[14] In practice, the IRSD is originally constructed for SA2 regions, which is a more granular geographic unit defined by the ABS. We construct an aggregated-version of the IRSD for SA4 regions by computing a population-weighted average of the IRSD for all SA2 regions that fall within each SA4 region. We provide additional details in the Supplement.

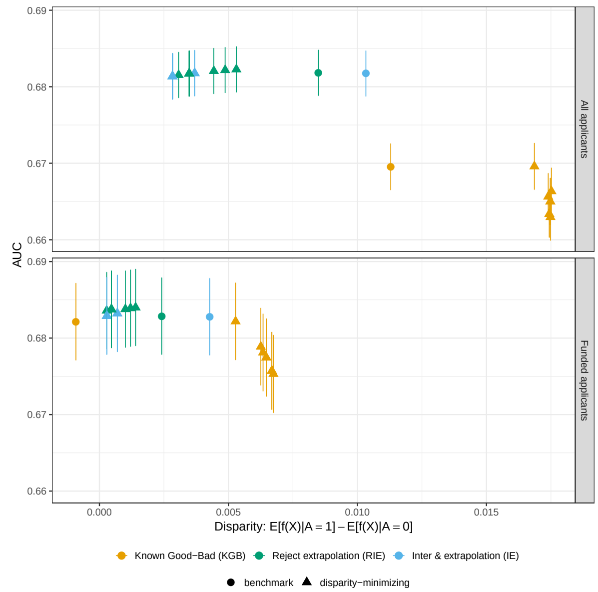


Figure 1: Area under the ROC curve (AUC) with respect to the synthetic outcome $\tilde{Y}_i$ against disparity in the average risk prediction for the disadvantaged ($A = 1$) vs advantaged ($A = 0$) groups. Our method is able to reduce disparities for the RIE and IE approaches with negligible loss in performance (first row). Evaluation on only funded (known) applicants (second row) is misleading in suggesting that the KGB models have comparable AUC performance to the RIE and IE models and in underestimating disparities for all models. Properly accounting for selective labels in the learning phase (RIE and IE) and in evaluation (first row) is critical for building models that perform well and reduce disparities on the full population. Error bars show the 95% bootstrap confidence intervals. See § 6 for further details.

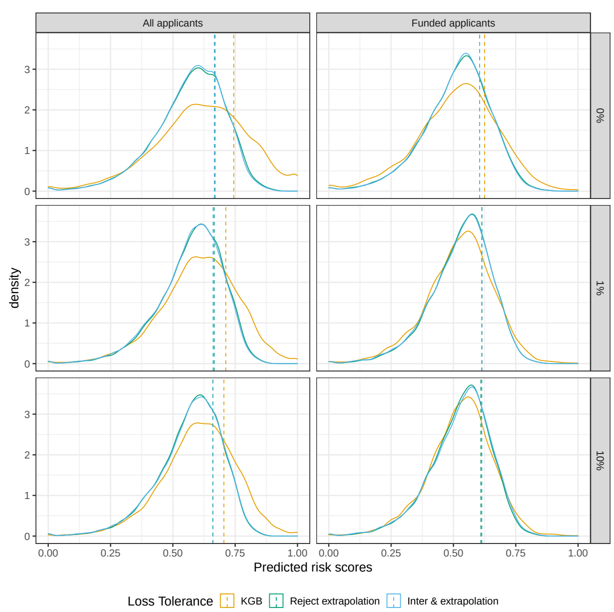
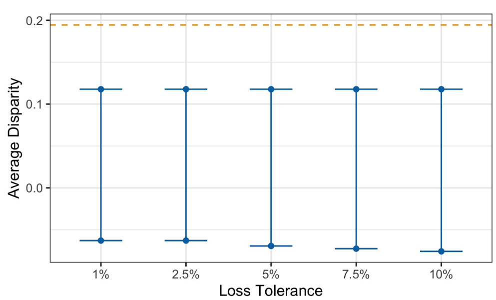


Figure 2: Distributions of predicted risk of defaulted for disadvantaged group $A = 1$ for the KGB, RIE and IE methods. The first row shows the benchmark model risk scores and subsequent rows show the risks for methods learned from our algorithm with loss tolerances of 1% and 10%. The left column shows risk scores on all applicants from the disadvantaged group. The right column shows only funded applicants from the disadvantaged group. Dashed line indicates the 75-percentile score. The RIE and IE methods predict lower rates of default for the disadvantaged group than the KGB method. Increasing the loss tolerance reduces the predicted risks for the KGB method. The densities for the funded applicants (right column) underestimate the differences in risk scores across the KGB, RIE, and IE methods (compare to left column). See § 6 for further details.

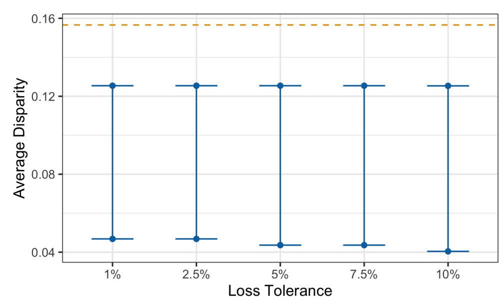
7 Application: Recidivism Risk Prediction

We apply our methods to explore the range of disparities over the set of good models in a recidivism risk prediction exercise applied to ProPublica's COMPAS recidivism data [49]. The goal of the following experiment is to illustrate how an auditor may use our methods to characterize the range of predictive disparities that may be generated over the set of good models, and examine whether the COMPAS tool generates larger disparities than any competing good model. As discussed earlier, this type of analysis is a crucial step to assessing legal claims of disparate impact.

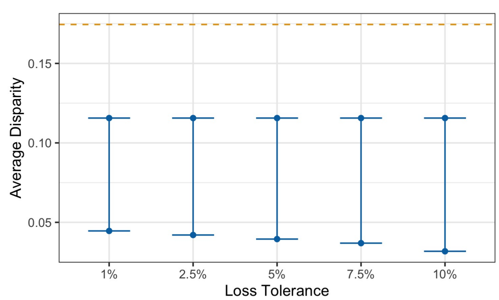
We focus on the range of predictive disparities between black $(A_i = 1)$ and white $(A_i = 0)$ defendants that could be generated by a recidivism risk score that is constructed using logistic regression on a quadratic polynomial of the defendant's age and number of prior offenses.[15] We set the loss tolerance ϵ such that (3) constructs the possible predictive range of disparities over all models that achieve a logistic regression loss within $\{1\%, 2.5\%, 5\%, 7.5\%, 10\%\}$ of the COMPAS score. We analyze the range of predictive disparities for statistical parity, balance for the positive class, and balance for the negative class. We split the data 50%-50% into a train set and test set, reporting the results on the test set. In each exercise, we ran Algorithm 1 on the train set for at most 500 iterations and defined the grid to be $\{1/40, 2/40, \ldots, 1\}$.



(a) Statistical Parity



(b) Balance for the Positive Class



(c) Balance for the Negative Class

Figure 3: Range of average disparities between black defendants $(A = 1)$ and white defendants $(A = 0)$ over the test set as the loss tolerance varies. The loss tolerance is calibrated using the COMPAS score as the benchmark model. The COMPAS score generates larger predictive disparities between black and white defendants than any model in the set of good models. See § 7 for further details.

Figure 3 summarizes the range of predictive disparities that could be generated over the set of good models within a neighborhood of the COMPAS score. The blue error bars plot the range of predictive disparities that are associated with the linear program reduction [§ 4.2] and the orange dashed line plots the predictive disparity associated with the COMPAS score itself. There are several interesting features worth highlighting. First, for each predictive disparity measure, we find that the range of predictive disparities between black and white defendants increases as the loss tolerance grows. As the loss tolerance grows, we are imposing a "weaker" definition of a good model in (3). Second, for balance for the positive class and balance for the negative class, the minimal predictive disparity over the set of good

[15] We restrict the analysis to black and white defendants.

models between black and white defendants is positive for all choices of the loss tolerance parameter. Provided that the decision-maker is required to search for prediction functions within a neighborhood of the COMPAS score, any resulting model produces positive predictive disparities black and white defendants on these measures. Finally, the predictive disparity of the COMPAS score is always strictly larger than the maximal predictive disparity that could be generated over the set of good models, no matter the definition of predictive disparity. For statistical parity, balance for the positive class and balance for the negative class, the COMPAS score produces a larger predictive disparity between black and white defendants than the worst-case risk score constructed using logistic regression on a quadratic polynomial of the defendant's age and number of prior offenses.

Additional Experiments: In the Supplement, we document performance on the train set and examine the range of bounded group loss predictive disparities. We also analyze the range of predictive disparities between young and older defendants over the set of good models in the ProPublica COMPAS data. Finally, we analyze the UCI adult income dataset, which is another common benchmark.

8 Conclusion

In this paper, we provided a generic method for characterizing predictive disparities over the set of good models and finding the absolute disparity minimizing model over the set of good models, which we defined as the set of prediction functions in a chosen model class that achieve some average loss. In settings with selective labels, our method can be used to characterize fairness over the full population. We establish guarantees on the generalization error and predictive disparities of our approach. Our empirical analysis illustrated how to use our method for two use cases: (1) finding a more equitable model with performance comparable to the benchmark model; and (2) audits for disparate impact. In many settings the set of good models is a rich class, whereby models differ substantially in terms of their fairness properties. Our approach leverages this to reduce disparities without compromising on overall performance.

Acknowledgements

We are especially 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 thank Sendhil Mullainathan for valuable comments and feedback. Rambachan gratefully acknowledges financial support from the NSF Graduate Research Fellowship under Grant DGE1745303.

References

[1] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 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*, page 1721–1730, 2015.

[2] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. volume 56 of *Proceedings of Machine Learning Research*, pages 301–318, 2016.

[3] Sendhil Mullainathan and Ziad Obermeyer. Who is tested for heart attack and who should be: predicting patient risk and physician error. Technical report, NBER Working Paper Series, Working Paper No. 26168, 2020.

[4] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. volume 81 of *Proceedings of Machine Learning Research*, pages 134–148, 2018.

[5] Himabindu Lakkaraju and Cynthia Rudin. Learning Cost-Effective and Interpretable Treatment Regimes. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 166–175, 2017.

[6] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2018.

[7] Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767 – 2787, 2010.

[8] Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. Predictably unequal? the effects of machine learning on credit markets. Technical report, 2020.

[9] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, page 469–481, 2020.

[10] Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–215, 2001.

[11] Alexandra Chouldechova and Max G'Sell. Fairer and more accurate, but for whom? Technical report, In Proceedings of the 2017 FAT/ML Workshop, 2017.

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

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

[14] Solon Barocas and Andrew Selbst. Big data's disparate impact. *California Law Review*, 104:671–732, 2016.

[15] Talia Gillis. False dreams of algorithmic fairness: The case of credit pricing. 2019.

[16] Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 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*, pages 275–284, 2017.

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

[18] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Technical report, arXiv preprint arXiv:1801.01489, 2019.

[19] Jiayun Dong and Cynthia Rudin. Variable importance clouds: A way to explore variable importance for the set of good models. Technical report, arXiv preprint arXiv:1901.03209, 2020.

[20] Lesia Semenova, Cynthia Rudin, and Ronald Parr. 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.

[21] Charles T. Marx, Flavio du Pin Calmon, and Berk Ustun. Predictive multiplicity in classification. Technical report, arXiv preprint arXiv:1909.06677, 2019.

[22] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, pages 60–69, 2018.

[23] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, 2019.

[24] Cynthia Dwork, Toniann Pitassi Moritz Hardt, Omer Reingold, and Richard Zemel. Fairness through awareness. In *ITCS '12: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

[25] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. pages 325–333, 2013.

[26] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 3323–3331, 2016.

[27] Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 107–118. 2018.

[28] Michele Donini, Luca Oneto, Shai Ben-David, John R Shawe-Taylor, and Massimiliano A. Pontil. Empirical risk minimization under fairness constraints. In *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 2796–2806, 2018.

[29] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

[30] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 5680–5689. 2017.

[31] Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, page 247–254, 2019.

[32] Ha-Thu Nguyen et al. Reject inference in application scorecards: evidence from france. Technical report, University of Paris Nanterre, EconomiX, 2016.

[33] Guoping Zeng and Qi Zhao. A rule of thumb for reject inference in credit scoring. *Math. Finance Lett.*, 2014:Article–ID, 2014.

[34] Amanda Coston, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation and fairness. In *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 582–593, 2020.

[35] Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. *arXiv preprint arXiv:1806.02887*, 2018.

[36] Zhiyong Li, Xinyi Hu, Ke Li, Fanyin Zhou, and Feng Shen. Inferring the outcomes of rejected loans: An application of semisupervised clustering. *Journal of the Royal Statistical Society: Series A*, 183(2):631–654, 2020.

[37] Rogelio A Mancisidor, Michael Kampffmeyer, Kjersti Aas, and Robert Jenssen. Deep generative models for reject inference in credit scoring. *Knowledge-Based Systems*, page 105758, 2020.

[38] David J Hand and William E Henley. Can reject inference ever work? *IMA Journal of Management Mathematics*, 5(1):45–55, 1993.

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

[40] Yahav Bechavod, Katrina Ligett, Aaron Roth, Bo Waggoner, and Steven Z. Wu. Equal opportunity in online classification with partial feedback. In *Advances in Neural Information Processing Systems 32*, pages 8974–8984. 2019.

[41] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

[42] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

[43] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. pages 797–806, 2017.

[44] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. Technical report, arXiv Working Paper, arXiv:1811.07867, 2019.

[45] Andrew Cotter, Heinrich Jiang, and Karthik Sridharan. Two-player games for efficient non-convex constrained optimization. pages 300–332, 2019.

[46] Guido W Imbens and Donald B Rubin. *Causal Inference for statistics, social and biomedical sciences: an introduction*. Cambridge University Press, Cambridge, United Kingdom, 2015.

[47] James J. Heckman. Varieties of selection bias. 80(2):313–318, 1990.

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

[49] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. there's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016.

[50] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

A Additional Theoretical Results

A.1 Computing the Absolute Predictive Disparity Minimizing Model

In this section, we show how the reductions approach may be extended to compute the prediction function that minimizes the absolute predictive disparity over the set of good models (5).

We again consider randomized prediction functions and construct a solution to

$$\min_{Q \in \Delta(\mathcal{F})} |\text{disparity}(Q)| \text{ subject to } \text{loss}(Q) \leq \epsilon. \tag{24}$$

Through the same discretization argument, this problem may be reduced to a constrained classification problem over the set of threshold classifiers

$$\min_{Q_h \in \Delta(\mathcal{H})} |\text{disparity}(Q_h)| \text{ subject to } \text{cost}(Q_h) \leq \epsilon - c_0. \tag{25}$$

To further deal with the absolute value operator in the objective function, we introduce a slack variable ξ and define the equivalent problem over both $Q_h \in \Delta(\mathcal{H}), \xi \in \mathbb{R}$

$$\begin{aligned} \min_{\xi, Q_h \in \Delta(\mathcal{H})} \; & \xi \\ \text{subject to } & \text{disparity}(Q_h) - \xi \leq 0, \\ & -\text{disparity}(Q_h) - \xi \leq 0, \\ & \text{cost}(Q_h) \leq \epsilon - c_0. \end{aligned} \tag{26}$$

We construct solutions to the empirical analogue of (26).

A.1.1 The Algorithm

Solving the empirical analogue of (26) is equivalent to finding the saddle point

$$\min_{Q_h \in \Delta(\mathcal{H}), \xi \in [0, B_\xi]} \max_{\|\lambda\| \leq B_\lambda} L(\xi, Q_h, \lambda), \tag{27}$$

where the Lagrangian is now defined as

$$\begin{aligned} L(\xi, Q_h, \lambda) =& \xi + \lambda_+ \left(\hat{\text{disparity}}(Q_h) - \xi \right) \\ & + \lambda_- \left(-\hat{\text{disparity}}(Q_h) - \xi \right) + \lambda_{\text{cost}} \left(\widehat{\text{cost}}(Q_h) - \hat{\epsilon} \right), \end{aligned} \tag{28}$$

$\lambda = (\lambda_+, \lambda_-, \lambda_{\text{cost}})$ and B_ξ is a bound on the slack variable.[16] We search for the saddle point by treating it as the equilibrium of a two-player zero-sum game in which one player chooses (ξ, Q_h) and the other chooses λ.

Algorithm 4 computes a ν-approximate saddle point of $L(\xi, Q_h, \lambda)$. The best-response of the λ-player sets the Lagrange multiplier associated with the maximally violated constraint equal to B_λ. Otherwise, she sets all Lagrange multipliers to zero if all constraints are satisfied. In order to analyze the best-response of the (ξ, Q_h)-player, we first rewrite the Lagrangian as

$$\begin{aligned} L(\xi, Q_h, \lambda) =& (1 - \lambda_+ - \lambda_-)\xi \\ & + (\lambda_+ - \lambda_-)\widehat{\text{disparity}}(Q_h) + \lambda_{\text{cost}}(\widehat{\text{cost}}(Q_h) - \hat{\epsilon}). \end{aligned} \tag{29}$$

For a fixed value of λ, minimizing $L(\xi, Q_h, \lambda)$ over (ξ, Q_h) jointly is equivalent to separately minimizing the first term involving ξ and the remaining terms involving Q_h. To minimize $(1 - \lambda_+ - \lambda_-)\xi$, the best-response is to set $\xi = B_\xi$ if $1 - \lambda_+ - \lambda_- < 0$, and set $\xi = 0$ otherwise. Minimizing

$$(\lambda_+ - \lambda_-)\widehat{\text{disparity}}(Q_h) + \lambda_{\text{cost}}(\widehat{\text{cost}}(Q_h) - \hat{\epsilon}) \tag{30}$$

over Q_h can be achieved through a reduction to cost-sensitive classification since minimizing the previous display is equivalent to minimizing

$$\hat{\mathbb{E}}\left[\mathbb{E}_{Z_\alpha}\left[c_\lambda(\underline{Y}_i, A_i, Z_\alpha) h_f(X_i, Z_\alpha)\right]\right], \tag{31}$$

[16]Since the absolute predictive disparity is bounded by one, we define $B_\xi = 1$ in practice.

where now $c_\lambda(\underline{Y}_i, A_i, Z_\alpha) := (\lambda_+ - \lambda_-)\left(\frac{\beta_0}{\hat{p}_0} 1\{\mathcal{E}_0\} + \frac{\beta_1}{\hat{p}_1} 1\{\mathcal{E}_1\}\right) + \lambda_{\text{cost}} c(\underline{Y}_i, Z_\alpha)$.

Algorithm 4: Algorithm for finding the absolute predictive disparity minimizing model among the set of good models

Input: Training data $\{(X_i, Y_i, A_i)\}_{i=1}^n$,
Parameters β_0, β_1, Events $\mathcal{E}_0, \mathcal{E}_1$, and empirical loss tolerance $\hat{\epsilon}$
Bounds B_λ, B_ξ, accuracy ν and learning rate η
Result: ν-approximate saddle point $(\hat{\xi}, \hat{Q}, \hat{\lambda})$
Set $\theta_1 = 0 \in \mathbb{R}^3$;
for $t = 1, 2, \ldots$ **do**
  Set $\lambda_{t,k} = B_\lambda \frac{\exp(\theta_{t,k})}{1 + \sum_{k'} \exp(\theta_{t,k'})}$ for all $k = \{\text{cost}, +, -\}$;
  $h_t \leftarrow \text{Best}_h(\lambda_t)$, $\xi_t \leftarrow \text{Best}_\xi(\lambda_t)$;
  $\hat{Q}_{h,t} \leftarrow \frac{1}{t}\sum_{s=1}^t h_s$, $\hat{\xi}_t \leftarrow \frac{1}{t}\sum_{s=1}^t \xi_t$;
  $\bar{L} \leftarrow L(\hat{\xi}_t, \hat{Q}_t, \text{Best}_\lambda(\hat{\xi}_t.\hat{Q}_t)$;
  $\hat{\lambda}_t \leftarrow \frac{1}{t}\sum_{s=1}^t \lambda_s$, $\underline{L} \leftarrow L(\text{Best}_\xi(\lambda_t), \text{Best}_h(\hat{\lambda}_t), \hat{\lambda}_t)$;
  $\nu_t \leftarrow \max\left\{L(\hat{\xi}_t, \hat{Q}_t, \hat{\lambda}_t) - \underline{L}, \bar{L} - L(\hat{\xi}_t, \hat{Q}_t, \hat{\lambda}_t)\right\}$;
  if $\nu_t \le \nu$ **then**
    if $\widehat{\text{cost}}(\hat{Q}) \le \hat{\epsilon} + \frac{B_\xi + 2\nu}{B_\lambda}$ **then**
      return $(\hat{\xi}_t, \hat{Q}_t, \hat{\lambda}_t)$;
    else
      return $null$;
    end
  end
  Set $\theta_{t+1} = \theta_t + \eta \begin{pmatrix} \widehat{\text{disparity}}(h_t) - \xi_t \\ -\widehat{\text{disparity}}(h_t) - \xi_t \\ \widehat{\text{cost}}(h_t) - \hat{\epsilon} \end{pmatrix}$;
end

We use an analogous linear program reduction (§ 4.2) to shrink the support of the solution returned by Algorithm 4.

A.1.2 Error Analysis

We analyze the suboptimality of the solution returned by Algorithm 4.

Theorem 3. *Suppose Assumption 1 holds for $C' \ge 2C + 2 + \sqrt{2\ln(8N/\delta)}$ and $C'' \ge \sqrt{\frac{-\log(\delta/8)}{2}}$.*

Then, Algorithm 4 with $\nu \propto n^{-\phi}, B_\lambda \propto n^\phi, N \propto n^\phi$ terminates in at most $O(n^{4\phi})$ iterations. It 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. *$\hat{Q}_h \ne null$ and for any $\tilde{Q}$ that is feasible in (24)*

$$\text{loss}(\hat{Q}_h) \le \epsilon + \tilde{O}(n^{-\phi}) \tag{32}$$

$$\left|\text{disparity}(\hat{Q}_h)\right| \le \left|\text{disparity}(\tilde{Q})\right| + \tilde{O}(n_0^{-\phi}) + \tilde{O}(n_1^{-\phi}). \tag{33}$$

2. *$\hat{Q}_h = null$ and (24) is infeasible.*

We next provide an oracle result for the absolute disparity minimizing algorithm under selective labels.

Theorem 4 (Selective Labels for Algorithm 4)**.** *Suppose Assumption 2 holds and Algorithm 1 is given as input the modified training data $\{(X_i, A_i, \mu(X_i)\}_{i=1}^n$.*

Under the same conditions as Theorem 3, Algorithm 4 terminates in at most $O(n^{4\phi})$ iterations. It 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. *$\hat{Q}_h \neq null$ and for any $\tilde{Q}$ that is feasible in (24)*

$$\mathrm{loss}_\mu(\hat{Q}_h) \leq \epsilon + \tilde{O}(n^{-\phi}) \tag{34}$$

$$\left|\mathrm{disparity}(\hat{Q}_h)\right| \leq \left|\mathrm{disparity}(\tilde{Q})\right| + \tilde{O}(n_0^{-\phi}) + \tilde{O}(n_1^{-\phi}). \tag{35}$$

2. *$\hat{Q}_h = null$ and (24) is infeasible.*

Proof. The proof is omitted since the analogous steps are given in proofs of Theorems 2-3. □

A.2 Bounded Group Loss Disparity

Bounded group loss is a common notion of predictive fairness that examines the variation in average loss across values of the protected or sensitive attribute. It is commonly used in fair classification and fair regression procedures to ensure that the prediction function achieves some minimal threshold of predictive performance across all values of the attribute [23].

We define a bounded group loss disparity to be the difference in average loss across values of the attribute

$$\mathrm{disparity}(f) = \mathbb{E}\left[l(Y_i, f(X_i)) \mid A_i = 1\right] - \mathbb{E}\left[l(Y_i, f(X_i)) \mid A_i = 0\right].$$

This choice of predictive disparity measure is convenient as it allows us to drastically simplify our algorithm by skipping the discretization step entirely and reducing the problem to an instance of weighted loss minimization. [23] apply the same idea in their analysis of fair regression under bounded group loss.

Take, for example, the problem of finding the range of bounded group loss disparities that are possible over the set of good models. Letting $\mathrm{loss}(f \mid A_i = a) := \mathbb{E}\left[l(Y_i, f(X_i) \mid A_i = a\right]$ and $\mathrm{loss}(Q \mid A_i = a) := \sum_{f \in \mathcal{F}} Q(f)\,\mathrm{loss}(f \mid A_i = a)$, we wish to solve

$$\min_{Q \in \Delta(\mathcal{F})} \mathrm{loss}(Q \mid A_i = 1) - \mathrm{loss}(f \mid A_i = 0), \text{ s.t. } \mathrm{loss}(Q) \leq \epsilon.$$

The sample version of this problem is

$$\min_{Q \in \Delta(\mathcal{F})} \widehat{\mathrm{loss}}(Q \mid A_i = 1) - \widehat{\mathrm{loss}}(f \mid A_i = 0) \tag{36}$$
$$\text{subject to } \widehat{\mathrm{loss}}(Q) \leq \epsilon.$$

We solve the sample problem by finding a saddle point of the associated Lagrangian

$$L(Q, \lambda) = \widehat{\mathrm{loss}}(Q \mid A_i = 1) - \widehat{\mathrm{loss}}(f \mid A_i = 0) + \lambda(\widehat{\mathrm{loss}}(Q) - \epsilon).$$

We compute a ν-approximate saddle point by treating it as a zero-sum game between a Q-player and a λ-player. The best response of the λ-player is the same as before: if the constraint $\hat{\mathrm{loss}}(Q) - \epsilon$ is violated, she sets $\lambda = B_\lambda$, and otherwise she sets $\lambda = 0$. The best-response of the Q-player may reduced to an instance of weighted loss minimization since

$$\begin{aligned}
&\widehat{\mathrm{loss}}(f|\mathcal{E}_0) - \widehat{\mathrm{loss}}(f|\mathcal{E}_1) + \lambda(\widehat{\mathrm{loss}}(f) - \epsilon) \\
&= \hat{\mathbb{E}}\left[\left(\frac{1}{\hat{p}_0} 1\{\mathcal{E}_0\} - \frac{1}{\hat{p}_1} 1\{\mathcal{E}_1\} + \lambda\right) l(Y_i, f(X_i))\right]
\end{aligned}$$

Therefore, defining the weights $W_i = \frac{1}{\hat{p}_0} 1\{\mathcal{E}_0\} - \frac{1}{\hat{p}_1} 1\{\mathcal{E}_1\} + \lambda$, we see that minimizing $L(h, \lambda)$ is equivalent to solving an instance of weighted loss minimization. From here, the analysis is the same as in § 4.1. Algorithm 5 formally states the procedure for finding the range of bounded group loss disparities. We may analogously extend Algorithm 4 to find the absolute bounded group loss minimizing model among the set of good models.

Algorithm 5: Algorithm for finding the bounded group loss disparity minimizing model over the set of good models

Input: Training data $\{(X_i, Y_i, A_i)\}_{i=1}^n$,
Parameters β_0, β_1, Events $\mathcal{E}_0, \mathcal{E}_1$, and loss tolerance $\hat{\epsilon}$
Bound B_λ, accuracy ν and learning rate η
Result: ν-approximate saddle point $(\hat{Q}_h, \hat{\lambda})$
Set $\theta_1 = 0 \in \mathbb{R}$;
for $t = 1, 2, \ldots$ **do**
  Set $\lambda_t = B_\lambda \frac{\exp(\theta_t)}{1+\exp(\theta_t)}$;
  $f_t \leftarrow \mathrm{Best}_f(\lambda_t)$;
  $\hat{Q}_t \leftarrow \frac{1}{t}\sum_{s=1}^t f_s, \quad \bar{L} \leftarrow L(\hat{Q}_t, \mathrm{Best}_\lambda(\hat{Q}_t))$;
  $\hat{\lambda}_t \leftarrow \frac{1}{t}\sum_{s=1}^t \lambda_s, \quad \underline{L} \leftarrow L(\mathrm{Best}_f(\hat{\lambda}_t), \hat{\lambda}_t)$;
  $\nu_t \leftarrow \max\left\{L(\hat{Q}_t, \hat{\lambda}_t) - \underline{L}, \bar{L} - L(\hat{Q}_t, \hat{\lambda}_t)\right\}$;
  if $\nu_t \leq \nu$ **then**
    if $\widehat{\mathrm{loss}}(\hat{Q}_t) \leq \hat{\epsilon} + \frac{|\beta_0|+|\beta_1|+2\nu}{B_\lambda}$ **then**
      return $(\hat{Q}_t, \hat{\lambda}_t)$;
    else
      return *null*
    end
  end
  Set $\theta_{t+1} = \theta_t + \eta\left(\widehat{\mathrm{loss}}(f_t) - \hat{\epsilon}\right)$;
end

B Proofs of Main Results

Proof of Lemma 1

The claim follows directly from Lemma 1 in [23]. We provide the proof for completeness. We first use the telescoping trick to obtain

$$\begin{aligned} l_\alpha(y, u) &= l(\underline{y}, [u]_\alpha + \frac{\alpha}{2}) \\ &= l(\underline{y}, \frac{\alpha}{2}) + \sum_{z_\alpha \in \mathcal{Z}_\alpha} \left[l(\underline{y}, z_\alpha + \frac{\alpha}{2}) - l(\underline{y}, z_\alpha - \frac{\alpha}{2}) \right] 1\{u \geq z_\alpha\}. \end{aligned}$$

Plugging in $u = \underline{f}(x)$ and using the fact that for $z_\alpha \in \mathcal{Z}_\alpha$, we have that $1\{\underline{f}(x) \geq z_\alpha\} = 1\{f(x) \geq z_\alpha\} = h_f(x, z_\alpha)$, we obtain

$$l_\alpha(y, \underline{f}(x)) = l(\underline{y}, \frac{\alpha}{2}) + \frac{1}{N} \sum_{z_\alpha \in \mathcal{Z}_\alpha} c(\underline{y}, z_\alpha) h_f(x, z_\alpha).$$

The constant c_0 equals $\mathbb{E}\left[l(\underline{Y}, \frac{\alpha}{2})\right]$. □

Proof of Lemma 2

Fix $f \in \mathcal{F}$. For $x \in \mathcal{X}$ and $z_\alpha \in \mathcal{Z}_\alpha$

$$h_f(x, z_\alpha) = 1\{f(x) \geq z_\alpha\} = 1\{\underline{f}(x) \geq z_\alpha\},$$

Therefore,

$$\mathbb{E}_{Z_\alpha}\left[h_f(x, Z_\alpha)\right] = \mathbb{E}_{Z_\alpha}\left[1\{\underline{f}(x) \geq Z_\alpha\}\right] = \underline{f}(x),$$

and for any $a \in \{0, 1\}$,

$$\begin{aligned}
&|\mathbb{E}[h_f(X, Z_\alpha)|\mathcal{E}_a] - \mathbb{E}[f(X)|\mathcal{E}_a]| \\
&= |\mathbb{E}[\mathbb{E}_{Z_\alpha}[h_f(X, Z_\alpha)] - f(X)|\mathcal{E}_a]| \\
&= |\mathbb{E}[\underline{f}(X) - f(X)|\mathcal{E}_a]| \leq \alpha
\end{aligned}$$

where the first equality uses iterated expectations plus the fact that Z_α is independent of (X, A, Y) and the final equality follows by the definition of $\underline{f}(X)$. The claim is immediate after noticing $\text{disparity}(h_f) - \text{disparity}(f)$ equals

$$\beta_0 \left(\mathbb{E}[h_f(X, Z_\alpha) - f(X)|\mathcal{E}_0]\right) + \beta_1 \left(\mathbb{E}[h_f(X, Z_\alpha) - f(X)|\mathcal{E}_1]\right)$$

and applying the triangle inequality. □

Proof of Theorem 1

The claim about the iteration complexity of Algorithm 1 follows immediately from Lemma 3, substituting in the stated choices of ν and B.

The proof strategy for the remaining claims follows the proof of Theorems 2-3 in [23]. We consider two cases.

Case 1: There is a feasible solution Q^* to the population problem (6) Using Lemmas 5-6, the ν-approximate saddle point $\hat{Q}_h$ satisfies

$$\widehat{\text{disparity}}(\hat{Q}_h) \leq \widehat{\text{disparity}}(Q_h) + 2\nu \tag{37}$$

$$\widehat{\text{cost}}(\hat{Q}_h) \leq \hat{\epsilon} + \frac{|\beta_0| + |\beta_1| + 2\nu}{B} \tag{38}$$

for any distribution Q_h that is feasible in the empirical problem (11). This implies that Algorithm 1 returns $\hat{Q} \neq null$. We now show that the returned $\hat{Q}_h$ provides an approximate solution to the discretized population problem.

First, define

$$\widehat{\text{cost}}_z(h) := \hat{\mathbb{E}}[c(\underline{Y}_i, z)h(X_i, z)] \text{ and } \text{cost}_z(h) := \mathbb{E}[c(\underline{Y}_i, z)h(X_i, z)].$$

Since $c(\underline{Y}_i, z) \in [-1, 1]$, we invoke Lemma 8 with $S_i = c(\underline{Y}_i, z_i)$, $U_i = (X_i, z)$, $\mathcal{G} = \mathcal{H}$ and $\psi(s, t) = st$ to obtain that with probability at least $1 - \frac{\delta}{4}$ for all $z \in \mathcal{Z}_\alpha$ and $h \in \mathcal{H}$

$$\left|\widehat{\text{cost}}_z(h) - \text{cost}_z(h)\right| \leq 2R_n(\mathcal{H}) + \frac{2}{\sqrt{n}} + \sqrt{\frac{2\ln(8N/\delta)}{n}} = \tilde{O}(n^{-\phi}),$$

where the last equality follows by the bound on $R_n(\mathcal{H})$ in Assumption 1 and setting $N \propto n^\phi$. Averaging over $z \in \mathcal{Z}_\alpha$ and taking a convex combination of according to $Q_h \in \Delta(\mathcal{H})$ then delivers via Jensen's Inequality that with probability at least $1 - \delta/4$ for all $Q \in \Delta(\mathcal{H})$

$$\left|\widehat{\text{cost}}(Q_h) - \text{cost}(Q_h)\right| \leq \tilde{O}(n^{-\phi}). \tag{39}$$

Next, define

$$\begin{aligned}
\widehat{\text{disparity}}_z(h) &:= \beta_0 \hat{\mathbb{E}}[h(X_i, z)|\mathcal{E}_0] + \beta_1 \hat{\mathbb{E}}[h(X_i, z)|\mathcal{E}_1] \\
\text{disparity}_z(h) &:= \beta_0 \mathbb{E}[h(X_i, z)|\mathcal{E}_0] + \beta_1 \mathbb{E}[h(X_i, z)|\mathcal{E}_1],
\end{aligned}$$

where the difference can be expressed as

$$\begin{aligned}
&\widehat{\text{disparity}}_z(h) - \text{disparity}_z(h) = \\
&\beta_0 \left(\hat{\mathbb{E}}[h(X_i, z)|\mathcal{E}_0] - \mathbb{E}[h(X_i, z)|\mathcal{E}_0]\right) + \\
&\beta_1 \left(\hat{\mathbb{E}}[h(X_i, z)|\mathcal{E}_1] - \mathbb{E}[h(X_i, z)|\mathcal{E}_1]\right).
\end{aligned}$$

Therefore, by the triangle inequality,

$$\begin{aligned}&\left|\widehat{\text{disparity}}_z(h) - \text{disparity}_z(h)\right| \leq \\ &|\beta_0| \left|\hat{\mathbb{E}}\left[h(X_i, z)|\mathcal{E}_0\right] - \mathbb{E}\left[h(X_i, z)|\mathcal{E}_0\right]\right| + \\ &|\beta_1| \left|\hat{\mathbb{E}}\left[h(X_i, z)|\mathcal{E}_1\right] - \mathbb{E}\left[h(X_i, z)|\mathcal{E}_1\right]\right|.\end{aligned}$$

For each term on the right-hand side of the previous display, we invoke Lemma 8 applied to the data distribution conditional on $\mathcal{E}_0$ and $\mathcal{E}_1$. We set $S = 1$, $U = (X_i, z)$, $\mathcal{G} = \mathcal{H}$ and $\psi(s, t) = st$. With probability at least $1 - \frac{\delta}{4}$ for all $z \in \mathcal{Z}_\alpha$,

$$\begin{aligned}\left|\hat{\mathbb{E}}\left[h(X_i, z)|\mathcal{E}_0\right] - \mathbb{E}\left[h(X_i, z)|\mathcal{E}_0\right]\right| &\leq R_{n_0}(\mathcal{H}) + \frac{2}{\sqrt{n_0}} + \sqrt{\frac{2\ln(8N/\delta)}{n_0}} \\ \left|\hat{\mathbb{E}}\left[h(X_i, z)|\mathcal{E}_1\right] - \mathbb{E}\left[h(X_i, z)|\mathcal{E}_1\right]\right| &\leq R_{n_1}(\mathcal{H}) + \frac{2}{\sqrt{n_1}} + \sqrt{\frac{2\ln(8N/\delta)}{n_1}},\end{aligned}$$

Then, averaging over $z \in \mathcal{Z}_\alpha$ and taking a convex combination according to $Q_h \in \Delta(\mathcal{H})$ delivers via Jensen's Inequality that with probability at least $1 - \delta/4$ for all $Q \in \Delta(\mathcal{H})$

$$\left|\hat{\mathbb{E}}\left[Q_h|\mathcal{E}_0\right] - \mathbb{E}\left[Q_h|\mathcal{E}_0\right]\right| \leq R_{n_0}(\mathcal{H}) + \frac{2}{\sqrt{n_0}} + \sqrt{\frac{2\ln(8N/\delta)}{n_0}} \tag{40}$$

$$\left|\hat{\mathbb{E}}\left[Q_h|\mathcal{E}_1\right] - \mathbb{E}\left[Q_h|\mathcal{E}_1\right]\right| \leq R_{n_1}(\mathcal{H}) + \frac{2}{\sqrt{n_1}} + \sqrt{\frac{2\ln(8N/\delta)}{n_1}} \tag{41}$$

By the union bound, both inequalities hold with probability at least $1 - \delta/2$.

Finally, Hoeffding's Inequality implies that with probability at least $1 - \delta/4$,

$$|\hat{c}_0 - c_0| \leq \sqrt{\frac{-\log(\delta/8)}{2n}}. \tag{42}$$

From Lemma 7, we have that Algorithm 1 terminates and delivers a distribution $\hat{Q}_h$ that compares favorably against any feasible Q in the discretized sample problem (11). That is, for any such Q_h,

$$\widehat{\text{disparity}}(\hat{Q}_h) \leq \widehat{\text{disparity}}(Q_h) + O(n^{-\phi}) \tag{43}$$

$$\widehat{\text{cost}}(\hat{Q}_h) \leq \hat{\epsilon} + O(n^{-\phi}) \tag{44}$$

where we used the fact that $\nu \propto n^{-\phi}$ and $B \propto n^{\phi}$ by assumption. First, (39), (42), (44) imply

$$\text{cost}(\hat{Q}_h) \leq \hat{\epsilon} + \tilde{O}(n^{-\phi}) \leq \epsilon - c_0 + \tilde{O}(n^{-\phi}), \tag{45}$$

where we used that $\hat{\epsilon} = \epsilon - \hat{\mathbb{E}}[l(\underline{Y}, \frac{\alpha}{2})] + C'n^{-\phi} - C''n^{-1/2}$. by assumption. Second, the bounds in (40), (41) imply

$$\text{disparity}(\hat{Q}_h) \leq \text{disparity}(Q_h) + \tilde{O}(n_0^{-\beta}) + \tilde{O}(n_1^{-\phi}). \tag{46}$$

We assumed that Q_h was a feasible point in the discretized sample problem (11). Assuming that (39) holds implies that any feasible solution of the population problem is also feasible in the empirical problem due to how we have set C' and C''. Therefore, we have just shown in (45), (46) that $\hat{Q}_h$ is approximately feasible and approximately optimal in the discretized population problem (10). Our last step is to relate $\hat{Q}_h$ to the original problem over $\Delta(\mathcal{F})$ (6).

From Lemma 1 and (45), we observe that

$$\text{loss}_\alpha(\hat{Q}_h) \overset{(1)}{\leq} \epsilon + \tilde{O}(n^{-\phi}),$$
$$\text{loss}(\hat{Q}_h) \overset{(2)}{\leq} \epsilon + \tilde{O}(n^{-\phi}),$$

where (1) used Lemma 1 and we now view $\hat{Q}_h$ as a distribution of risk scores $f \in \mathcal{F}$, (2) used that $\text{loss}(Q) \leq \text{loss}_\alpha(Q) + \alpha$. Next, from Lemma 2 and (46), we observe that

$$\text{disparity}(\hat{Q}_h) \leq \text{disparity}(\tilde{Q}) + (|\beta_0| + |\beta_1|)\,\alpha + \tilde{O}(n_0^{-\phi}) + \tilde{O}(n_1^{-\phi}).$$

where $\hat{Q}_h$ is viewed as a distribution over risk scores $f \in \mathcal{F}$ and $\tilde{Q}$ is now any distribution over risk scores $f \in \mathcal{F}$ that is feasible in the fairness frontier problem. This proves the result for Case I.

Case II: There is no feasible solution to the population problem (6) This follows the proof of Case II in Theorem 3 of [23]. If the algorithm returns a ν-approximate saddle point $\hat{Q}_h$, then the theorem holds vacuously since there is no feasible $\tilde{Q}$. Similarly, if the algorithm returns $null$, then the theorem also holds. □

Proof of Theorem 2

Under oracle access to $\mu(x)$, the iteration complexity and bound on cost hold immediately from Theorems 1. The bound on disparity holds immediately for choices $\mathcal{E}_0, \mathcal{E}_1$ that depend on only A. For choices of $\mathcal{E}_0, \mathcal{E}_1$ that depends on Y_i, such as the qualified affirmative action fairness-enhancing intervention, we rely on Lemma 9. We first observe that under oracle access to $\mu(x)$, we can identify any disparity as

$$\frac{\beta_1 \mathbb{E}[f(X)g(\mu(X)) \mid A=1]}{\mathbb{E}[g(\mu(X)) \mid A=1]} - \frac{\beta_0 \mathbb{E}[f(X)g(\mu(X)) \mid A=0]}{\mathbb{E}[g(\mu(X)) \mid A=0]}, \tag{47}$$

where $g(x) = x$ for the balance for the positive class and qualified affirmative action criteria; $g(x) = (1-x)$ for balance for the negative class; and $g(x) = 1$ for the statistical parity and the affirmative action criteria (see proof of Lemma 9 below proof for an example). We define the shorthand

$$\begin{aligned}
\omega_1 &:= \mathbb{E}[f(X)g(\mu(X)) \mid A=1] \\
\bar{\omega}_1 &:= \mathbb{E}[g(\mu(X)) \mid A=1] \\
\omega_0 &= \mathbb{E}[f(X)g(\mu(X)) \mid A=0] \\
\bar{\omega}_0 &= \mathbb{E}[g(\mu(X)) \mid A=0]
\end{aligned}$$

and we use $\hat{\omega}_1$, $\hat{\bar{\omega}}_1$, $\hat{\omega}_0$, and $\hat{\bar{\omega}}_0$ to denote their empirical estimates. Lemma 9 gives the following bound on the empirical estimate of the disparity:

$$\begin{aligned}
&\mathbb{P}\left[\left|\frac{\beta_1\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\beta_0\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \left(\frac{\beta_1\omega_1}{\bar{\omega}_1} - \frac{\beta_0\omega_0}{\bar{\omega}_0}\right)\right| \geq \epsilon\right] \\
&\leq 4\exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_\wedge}{8\beta} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right] + 2\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^4}{64\beta^2\omega_\vee^2}\right] \\
&+ 2\exp\left[\frac{-n\bar{\omega}_\wedge^2}{4}\right]
\end{aligned}$$

where $\omega_\vee = \max(\omega_1, \omega_0)$, $\bar{\omega}_\wedge = \min(\bar{\omega}_1, \bar{\omega}_0)$ and $\beta = \max(|\beta_1|, |\beta_0|)$.

We now proceed to relax and simplify the bound. For $\epsilon \leq 4\frac{\beta\omega_\vee}{\bar{\omega}_\wedge}$, we have

$$2\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^4}{64\beta^2\omega_\vee^2}\right] \geq 2\exp\left[\frac{-n\bar{\omega}_\wedge^2}{4}\right]$$

Case 1: We first consider the likely case that $\bar{\omega}_\wedge \geq \omega_\vee$. Then we have

$$2\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^4}{64\beta^2\omega_\vee^2}\right] \leq 2\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^2}{64\beta^2}\right]$$

1a) If

$$\frac{\epsilon\bar{\omega}_\wedge}{8\beta} \geq 4R_n(\mathcal{G}) + \frac{2}{\sqrt{n}} \tag{48}$$

then

$$\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^2}{64\beta^2}\right] \leq \exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_\wedge}{8\beta} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right]$$

Then we have

$$\mathbb{P}\left[\left|\frac{\beta_1\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\beta_0\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \left(\frac{\beta_1\omega_1}{\bar{\omega}_1} - \frac{\beta_0\omega_0}{\bar{\omega}_0}\right)\right| \geq \epsilon\right] \tag{49}$$

$$\leq 8\exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_\wedge}{8\beta} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right] \tag{50}$$

Inverting this bound yields the following: with probability at least $1-\delta$,

$$\left|\frac{\beta_1\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\beta_0\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \left(\frac{\beta_1\omega_1}{\bar{\omega}_1} - \frac{\beta_0\omega_0}{\bar{\omega}_0}\right)\right| \leq \frac{8\beta}{\bar{\omega}_\wedge}\left(4R_n(\mathcal{G}) + \frac{2}{\sqrt{n}} + \sqrt{\frac{2}{n}\log\left(\frac{8}{\delta}\right)}\right)$$

1b:

$$\frac{\epsilon\bar{\omega}_\wedge}{8\beta} < 4R_n(\mathcal{G}) + \frac{2}{\sqrt{n}} \tag{51}$$

implies that

$$\left|\frac{\beta_1\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\beta_0\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \left(\frac{\beta_1\omega_1}{\bar{\omega}_1} - \frac{\beta_0\omega_0}{\bar{\omega}_0}\right)\right| \leq \frac{8\beta}{\bar{\omega}_\wedge}\left(4R_n(\mathcal{G}) + \frac{2}{\sqrt{n}}\right)$$

Case 2: We now consider the unlikely but plausible case that $\bar{\omega}_\wedge < \omega_\vee$. Then we have

$$\exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_\wedge}{8\beta} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right] \leq \exp\left[-\frac{n}{2}\left(\frac{\epsilon\omega_\vee}{8\beta} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right]$$

and

$$\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^4}{64\beta^2\omega_\vee^2}\right] \leq \exp\left[\frac{-n\epsilon^2\omega_\vee^2}{64\beta^2}\right]$$

We proceed with the same steps as in Case 1 to conclude that with probability at least $1-\delta$,

$$\left|\frac{\beta_1\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\beta_0\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \left(\frac{\beta_1\omega_1}{\bar{\omega}_1} - \frac{\beta_0\omega_0}{\bar{\omega}_0}\right)\right| \leq \frac{8\beta}{\bar{\omega}_\wedge}\left(4R_n(\mathcal{G}) + \frac{2}{\sqrt{n}} + \sqrt{\frac{2}{n}\log\left(\frac{8}{\delta}\right)}\right)$$

Applying our assumption that

$$R_n(\mathcal{H}) \leq Cn^{-\phi} \text{ and } \hat{\epsilon} = \epsilon - \hat{c}_0 + C'n^{-\phi} - C''n^{-1/2}.$$

for $\phi \leq 1/2$ and $C' \geq 2C + 2 + \sqrt{2\ln(8N/\delta)}$ and $C'' \geq \sqrt{\frac{-\log(\delta/8)}{2}}$, then

$$\text{disparity}(\hat{Q}_h) \leq \text{disparity}(\tilde{Q}) + \tilde{O}(n^{-\phi}), \tag{52}$$

which implies

$$\text{disparity}(\hat{Q}_h) \leq \text{disparity}(\tilde{Q}) + \tilde{O}(n_0^{-\phi}) + \tilde{O}(n_1^{-\phi}). \tag{53}$$

□

Proof of Theorem 3

The claim about the iteration complexity of Algorithm 4 follows from Lemma 10 after substituting in the stated choices of ν, B_λ. We consider two cases.

Case 1: There is a feasible solution $\tilde{Q}$ to the population problem (24) Using Lemmas 12-14, the ν-approximate saddle point $(\hat{\xi}, \hat{Q}_h)$ satisfies

$$\widehat{\text{disparity}}(\hat{Q}_h) - \hat{\xi} \leq \frac{B_\xi + 2\nu}{B_\lambda}, \tag{54}$$

$$-\widehat{\text{disparity}}(\hat{Q}_h) - \hat{\xi} \leq \frac{B_\xi + 2\nu}{B_\lambda} \tag{55}$$

$$\widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon}_{\text{cost}} \leq \frac{B_\xi + 2\nu}{B_\lambda} \tag{56}$$

for any (ξ, Q) that is feasible in the empirical problem. This implies that Algorithm 4 returns $\hat{Q} \neq null$. We will now show that the $(\hat{\xi}, \hat{Q})$ provides an approximate solution to the discretized population problem.

First, through the same argument as in the proof of Theorem 1, we obtain that with probability at least $1 - \delta/4$ for all $Q_h \in \Delta(\mathcal{H})$

$$\left|\widehat{\text{cost}}(Q_h) - \text{cost}(Q_h)\right| \leq \tilde{O}(n^{-\phi}). \tag{57}$$

Second, with probability at least $1 - \delta/2$ for all $Q \in \Delta(\mathcal{H})$,

$$\left|\hat{\mathbb{E}}[Q_h|\mathcal{E}_0] - \mathbb{E}[Q_h|\mathcal{E}_0]\right| \leq \tilde{O}(n_0^{-\phi}) \tag{58}$$

$$\left|\hat{\mathbb{E}}[Q_h|\mathcal{E}_1] - \mathbb{E}[Q_h|\mathcal{E}_1]\right| \leq \tilde{O}(n_1^{-\phi}). \tag{59}$$

Finally, Hoeffding's Inequality implies that with probability at least $1 - \delta/4$,

$$|\hat{c}_0 - c_0| \leq \sqrt{\frac{-\log(\delta/8)}{2n}}. \tag{60}$$

From Lemma 15, we have that Algorithm 4 terminates and delivers $(\hat{\xi}, \hat{Q}_h)$ that compares favorable with any feasible (ξ, Q_h) in the discretized sample problem. That is, for any such (ξ, Q_h),

$$\hat{\xi} \leq \xi + O(n^{-\phi}), \tag{61}$$

$$\widehat{\text{disparity}}(\hat{Q}_h) \leq \hat{\xi} + O(n^{-\phi}), \tag{62}$$

$$-\widehat{\text{disparity}}(\hat{Q}_h) \leq \hat{\xi} + O(n^{-\phi}) \tag{63}$$

$$\widehat{\text{cost}}(\hat{Q}_h) \leq \hat{\epsilon}_{\text{cost}} + O(n^{-\phi}) \tag{64}$$

Notice that (57), (60) and (64) imply that

$$\text{cost}(\hat{Q}_h) \leq \epsilon - c_0 + \tilde{O}(n^{-\phi}), \tag{65}$$

where we used that $\hat{\epsilon} = \epsilon - \hat{c}_0 + C'n^{-\phi} - C''n^{-\phi}$. For any feasible (ξ, Q_h), then $(|\text{disparity}(Q_h)|, Q_h)$ is also feasible. Then, combining (61)-(63) yields

$$\left|\text{disparity}(\hat{Q}_h)\right| \leq |\text{disparity}(Q_h)| + \tilde{O}(n^{-\phi}) \tag{66}$$

Second, notice that this implies that

$$\left|\text{disparity}(\hat{Q}_h)\right| \leq |\text{disparity}(Q_h)| + \tilde{O}(n_0^{-\phi}) + \tilde{O}(n_1^{-\phi}) \tag{67}$$

We assumed that (ξ, Q_h) were feasible in the discretized sample problem. Assuming that (57) holds implies that any feasible solution of the population problem is also feasible in the empirical problem due to how we set C', C''. Therefore, we have just shown that $(\hat{\xi}, \hat{Q}_h)$ are approximately optimal in the discretized population problem.

Then, following the proof of Theorem 1, we observe that $\text{loss}(\hat{Q}_h) \leq \epsilon + \tilde{O}(n^{-\phi})$, where we now interpret $\hat{Q}_h$ as a distribution over risk scores $f \in \mathcal{F}$. This proves the result for Case I.

Case II: There is no feasible solution to the population problem (24) This follows the proof of Case II of Theorem 3 in [23]. If the algorithm returns a ν-approximate saddle point $\hat{Q}_h$, then the theorem holds vacuously since there is no feasible $\tilde{Q}$. Similarly, if the algorithm returns $null$, then the theorem also holds. □

C Auxiliary Lemmas for Main Results

In this section, we state and prove a series of auxiliary lemmas that are used in the proofs of our main results in the text.

C.1 Auxiliary Lemmas for the Proof of Theorem 1

C.1.1 Iteration Complexity of Algorithm 1

Lemma 3. *Letting $\rho := \max_{h\in\mathcal{H}} |\widehat{\text{cost}}(h) - \hat{\epsilon}|$, Algorithm 1 satisfies the inequality*

$$\nu_t \leq \frac{B\log(2)}{\eta t} + \eta\rho^2 B.$$

For $\eta = \frac{\nu}{2\rho^2 B}$, Algorithm 1 will return a ν-approximate saddle point of L in at most $\frac{4\rho^2 B^2 \log(2)}{\nu^2}$. Since in our setting, $\rho \leq 1$, the iteration complexity of Algorithm 1 is $4B^2\log(2)/\nu^2$.

Proof. Follows immediately from the proof of iteration complexity in Theorem 3 of [23]. Since the cost is bounded on $[-1, 1]$ and $\widehat{\text{cost}}(h) - \hat{\epsilon} \leq \widehat{\text{cost}}(h) \leq 1$ for any $h \in \mathcal{H}$, we see that $\rho \leq 1$. □

C.1.2 Solution Quality for Algorithm 1

Let $\Lambda := \{\lambda \in \mathbb{R}_+ : \lambda \leq B\}$ denote the domain of λ. Throughout this section, we assume we are given a pair $(\hat{Q}_h, \hat{\lambda})$ that is a ν-approximate saddle point of the Lagrangian

$$L(\hat{Q}_h, \hat{\lambda}) \leq L(Q_h, \hat{\lambda}) + \nu \text{ for all } Q_h \in \Delta(\mathcal{H}),$$
$$L(\hat{Q}_h, \hat{\lambda}) \geq L(\hat{Q}_h, \lambda) - \nu \text{ for all } 0 \leq \lambda \leq B.$$

We extend Lemma 1, Lemma 2 and Lemma 3 of [22] to our setting.

Lemma 4. *The pair $(\hat{Q}_h, \hat{\lambda})$ satisfies*

$$\hat{\lambda}\left(\widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon}\right) \geq B\left(\widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon}\right)_+ - \nu,$$

where $(x)_+ = \max\{x, 0\}$.

Proof. We consider a dual variable λ that is defined as

$$\lambda = \begin{cases} 0 \text{ if } \widehat{\text{cost}}(\hat{Q}_h) \leq \hat{\epsilon} \\ B \text{ otherwise.} \end{cases}$$

From the ν-approximate optimality conditions,

$$\begin{aligned} \widehat{\text{disparity}}(\hat{Q}) + \hat{\lambda}\left(\widehat{\text{cost}}(\hat{Q}) - \hat{\epsilon}\right) &= L(\hat{Q}, \hat{\lambda}) \\ &\geq L(\hat{Q}, \lambda) - \nu \\ &= \widehat{\text{disparity}}(\hat{Q}) + \lambda\left(\widehat{\text{cost}}(Q) - \hat{\epsilon}\right), \end{aligned}$$

and the claim follows by our choice of λ. □

Lemma 5. *The distribution $\hat{Q}_h$ satisfies*

$$\widehat{\text{disparity}}(\hat{Q}_h) \leq \widehat{\text{disparity}}(Q_h) + 2\nu$$

for any Q_h satisfying the empirical constraint (i.e., any Q_h such that $\widehat{\text{cost}}(Q_h) \leq \hat{\epsilon}$).

Proof. Assume Q_h satisfies $\widehat{\text{cost}}(Q_h) \leq \hat{\epsilon}$. Since $\hat{\lambda} \geq 0$, we have that

$$L(Q_h, \hat{\lambda}) = \widehat{\text{disparity}}(Q_h) + \hat{\lambda}\left(\widehat{\text{cost}}(Q_h) - \hat{\epsilon}\right) \leq \widehat{\text{disparity}}(Q_h).$$

Moreover, the ν-approximate optimality conditions imply that $L(\hat{Q}_h, \hat{\lambda}) \leq L(Q_h, \hat{\lambda}) + \nu$. Together, these inequalities imply that

$$L(\hat{Q}_h, \hat{\lambda}) \leq \widehat{\text{disparity}}(Q_h) + \nu.$$

Next, we use Lemma 4 to construct a lower bound for $L(\hat{Q}_h, \hat{\lambda})$. We have that

$$\begin{aligned} L(\hat{Q}_h, \hat{\lambda}) &= \widehat{\text{disparity}}(\hat{Q}_h) + \hat{\lambda}\left(\widehat{\text{cost}}(Q_h) - \hat{\epsilon}'\right) \\ &\geq \widehat{\text{disparity}}(\hat{Q}_h) + B\left(\widehat{\text{cost}}(\hat{Q}) - \hat{\epsilon}'\right)_+ - \nu \\ &\geq \widehat{\text{disparity}}(\hat{Q}_h) - \nu. \end{aligned}$$

By combining the inequalities $L(\hat{Q}_h, \hat{\lambda}) \geq \widehat{\text{disparity}}(\hat{Q}_h) - \nu$ and $L(\hat{Q}_h, \hat{\lambda}) \leq \widehat{\text{disparity}}(Q_h) + \nu$, we arrive at the claim. □

Lemma 6. *Assume the empirical constraint $\widehat{\text{cost}}(Q_h) \leq \hat{\epsilon}$ is feasible. Then, the distribution $\hat{Q}_h$ approximately satisfies the empirical cost constraint with*

$$\widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon} \leq \frac{|\beta_0| + |\beta_1| + 2\nu}{B}.$$

Proof. Let Q_h satisfy $\widehat{\text{cost}}(Q_h) \leq \hat{\epsilon}$. Recall from the proof of Lemma 5, we showed that

$$\widehat{\text{disparity}}(\hat{Q}_h) + B\left(\widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon}\right)_+ - \nu \leq L(\hat{Q}_h, \hat{\lambda}) \leq \widehat{\text{disparity}}(Q_h) + \nu.$$

Therefore, we observe that

$$B\left(\widehat{\text{cost}}(Q_h) - \hat{\epsilon}\right) \leq \left(\widehat{\text{disparity}}(Q_h) - \widehat{\text{disparity}}(\hat{Q}_h)\right) + 2\nu.$$

Since we can bound $\widehat{\text{disparity}}(Q_h) - \widehat{\text{disparity}}(\hat{Q}_h)$ by $|\beta_0| + |\beta_1|$, the result follows. □

Lemma 7. *Suppose that Q_h is any feasible solution to (11). Then, the solution $\hat{Q}_h$ returned by Algorithm 1 satisfies*

$$\begin{aligned} \widehat{\text{disparity}}(\hat{Q}_h) &\leq \widehat{\text{disparity}}(Q_h) + 2\nu \\ \widehat{\text{cost}}(\hat{Q}_h) &\leq \hat{\epsilon} + \frac{|\beta_0| + |\beta_1| + 2\nu}{B}. \end{aligned}$$

Proof. This is an immediate consequence of Lemma 3, Lemma 5 and Lemma 6. If the algorithm returns $null$, then these inequalities are vacuously satisfied. □

C.1.3 Concentration Inequality

We restate Lemma 2 in [23], which provides a uniform concentration inequality on the convergence of a sample moment over a function class.

Let $\mathcal{G}$ be a class of functions $g\colon \mathcal{U} \to \mathbb{R}$ over some space $\mathcal{U}$. The *Rademacher complexity* of the function class $\mathcal{G}$ is defined as

$$R_n(\mathcal{G}) := \sup_{u_1,\ldots,u_n \in \mathcal{U}} \mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(u_i)\right|\right],$$

where the expectation is defined over the i.i.d. random variables $\sigma_1, \ldots, \sigma_n$ with $P(\sigma_i = 1) = P(\sigma_i = -1) = 1/2$.

Lemma 8 (Lemma 2 in [23])**.** *Let D be a distribution over a pair of random variables (S, U) taking values in $\mathcal{S} \times \mathcal{U}$. Let $\mathcal{G}$ be a class of functions $g\colon \mathcal{U} \to [0, 1]$, and let $\psi\colon \mathcal{S} \times [0, 1] \to [-1, 1]$ be a contraction in its second argument* *(i.e., for all $s \in \mathcal{S}$ and $t, t' \in [0,1]$, $|\psi(s,t) - \psi(s,t')| \le |t - t'|$). Then, with probability $1-\delta$, for all $g \in \mathcal{G}$,*

$$\left|\hat{\mathbb{E}}\left[\psi(S, g(U))\right] - \mathbb{E}\left[\psi(S, g(U))\right]\right| \le 4R_n(\mathcal{G}) + \frac{2}{\sqrt{n}} + \sqrt{\frac{2\ln(2/\delta)}{n}},$$

where the expectation is with respect to D and the empirical expectation is based on n i.i.d. draws from D. If ψ is linear in its second argument, then a tighter bound holds with $4R_n(\mathcal{G})$ replaced by $2R_n(\mathcal{G})$.

C.2 Auxiliary Lemmas for the Proof of Theorem 2

C.2.1 Concentration result for disparity under selective labels

Lemma 9.

$$\begin{aligned}
&\mathbb{P}\left[\left|\frac{\beta_1\hat{\omega}_1}{\hat{\bar{\omega}}_1} - \frac{\beta_0\hat{\omega}_0}{\hat{\bar{\omega}}_0} - \left(\frac{\beta_1\omega_1}{\bar{\omega}_1} - \frac{\beta_0\omega_0}{\bar{\omega}_0}\right)\right| \ge \epsilon\right] \\
&\le 4\exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_\wedge}{8\beta} - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right] + 2\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^4}{64\beta^2\omega_\vee^2}\right] \\
&+ 2\exp\left[\frac{-n\bar{\omega}_\wedge^2}{4}\right]
\end{aligned}$$

where $\omega_\vee = \max(\omega_1, \omega_0)$, $\bar{\omega}_\wedge = \min(\bar{\omega}_1, \bar{\omega}_0)$ and $\beta = \max(|\beta_1|, |\beta_0|)$

Proof. For exposition, we first show the steps for qualified affirmative action and then extend the result to the general disparity. We can rewrite the qualified affirmative action criterion as

$$\mathbb{E}[f(X)|Y=1, A=1] = \frac{\mathbb{E}[f(X)\mu(X)|A=1]}{\mathbb{E}[\mu(X)|A=1]} \tag{68}$$

where $\mu(x) := \mathbb{E}[Y \mid X = x]$.

$\mathbb{E}[f(X)|Y=1, A=1]$

$$= \frac{\mathbb{E}[f(X)1\{Y=1\}|A=1]}{P(Y=1|A=1)} \tag{69}$$

$$= \frac{\mathbb{E}[f(X)\mathbb{E}[1\{Y=1\}|X,A=1]|A=1]}{E[P(Y=1|X,A=1)|A=1]} \tag{70}$$

$$= \frac{\mathbb{E}[f(X)P(Y=1|X,A=1)|A=1]}{E[\mu(X)|A=1]} \tag{71}$$

$$= \frac{\mathbb{E}[f(X)\mu(X)|A=1]}{E[\mu(X)|A=1]} \tag{72}$$

Assuming access to the oracle μ function, we can estimate this on the full training data as

$$\frac{\hat{\mathbb{E}}[f(X)\mu(X, A=1)|A=1]}{\hat{\mathbb{E}}[\mu(X, A=1)|A=1]} \tag{73}$$

Next we will make use of Lemma 2 of [23], which we restate here for convenience. Under certain conditions on ϕ and g, with probability at least $1-\delta$

$$\left|\hat{\mathbb{E}}\left[\phi(S, g(U))\right] - \mathbb{E}\left[\phi(S, g(U))\right]\right| \le 4R_n(\mathcal{G}) + \frac{2}{\sqrt{n}} + \sqrt{\frac{2\ln(2/\delta)}{n}}$$

We invert the bound by setting $\epsilon = 4R_n(\mathcal{G}) + \frac{2}{\sqrt{n}} + \sqrt{\frac{2\ln(2/\delta)}{n}}$ and solving for δ to get

$$\delta = 2\exp\left[-\frac{n}{2}\left(\epsilon - 4R_n(\mathcal{G}) - \frac{2}{\sqrt{n}}\right)^2\right] \tag{74}$$

Now we can restate Lemma 2 of [23] as

$$\mathbb{P}\Big[\,\Big|\hat{\mathbb{E}}\left[\phi(S,g(U))\right]-\mathbb{E}\left[\phi(S,g(U))\right]\Big|>\epsilon\Big] \tag{75}$$

$$\leq 2\exp\left[-\frac{n}{2}\left(\epsilon-4R_n(\mathcal{G})-\frac{2}{\sqrt{n}}\right)^2\right]$$

Next we revisit the quantity that we want to bound:

$$\left|\frac{\omega}{\bar{\omega}}-\frac{\hat{\omega}}{\hat{\bar{\omega}}}\right| \tag{76}$$

where $\omega=\mathbb{E}[f(X)\mu(X,A=1)|A=1]$ and $\bar{\omega}=\mathbb{E}[\mu(X,A=1)|A=1]$ and correspondingly for $\hat{\omega}$ and $\hat{\bar{\omega}}$. We will rewrite Expression 76 as a ratio of differences. We have

$$\left|\frac{\hat{\omega}}{\hat{\bar{\omega}}}-\frac{\omega}{\bar{\omega}}\right| \qquad = \left|\tfrac{\hat{\omega}\bar{\omega}-\hat{\bar{\omega}}\omega}{\hat{\bar{\omega}}\bar{\omega}}\right| \tag{77}$$

$$= \left|\tfrac{\bar{\omega}(\hat{\omega}-\omega)-\omega(\hat{\bar{\omega}}-\bar{\omega})}{\bar{\omega}(\hat{\bar{\omega}}-\bar{\omega})+\bar{\omega}^2}\right| \tag{78}$$

$$\tag{79}$$

By triangle inequality and union bound, we have

$$\mathbb{P}\Big[\Big|\frac{\bar{\omega}(\hat{\omega}-\omega)-\omega(\hat{\bar{\omega}}-\bar{\omega})}{\bar{\omega}(\hat{\bar{\omega}}-\bar{\omega})+\bar{\omega}^2}\Big|\geq\frac{t}{\bar{\omega}^2/2}\Big]$$

$$<\mathbb{P}\Big[|\bar{\omega}(\hat{\omega}-\omega)|+|\omega(\hat{\bar{\omega}}-\bar{\omega})|\geq t\Big]+\mathbb{P}\Big[|(\hat{\bar{\omega}}-\bar{\omega})+\bar{\omega}^2|\leq\frac{\bar{\omega}^2}{2}\Big]$$

$$<\mathbb{P}\Big[|\bar{\omega}(\hat{\omega}-\omega)|\geq\frac{t}{2}\Big]+\mathbb{P}\Big[|\omega(\hat{\bar{\omega}}-\bar{\omega})|\geq\frac{t}{2}\Big]+\mathbb{P}\Big[|\bar{\omega}(\hat{\bar{\omega}}-\bar{\omega})+\bar{\omega}^2|\leq\frac{\bar{\omega}^2}{2}\Big]$$

Since $0\leq\mu(X,A=1)\leq 1$, we can use a Hoeffding bound for the quantity $|(\hat{\bar{\omega}}-\bar{\omega})|$. Note that $0\leq\omega\leq\bar{\omega}\leq 1$. Then applying Hoeffding's inequality gives us

$$\mathbb{P}\Big[\,\big|\omega(\hat{\bar{\omega}}-\bar{\omega})\big|\geq\frac{t}{2}\Big]\leq 2\exp\left[\frac{-nt^2}{4\omega^2}\right] \tag{80}$$

Next we bound the third term:

$$\mathbb{P}\Big[|\bar{\omega}(\hat{\bar{\omega}}-\bar{\omega})+\bar{\omega}^2|\leq\frac{\bar{\omega}^2}{2}\Big] \quad \leq\mathbb{P}\Big[|\bar{\omega}(\hat{\bar{\omega}}-\bar{\omega})|\geq\tfrac{\bar{\omega}^2}{2}\Big] \tag{81}$$

$$=\mathbb{P}\Big[|(\hat{\bar{\omega}}-\bar{\omega})|\geq\tfrac{\bar{\omega}}{2}\Big] \tag{82}$$

$$\leq 2\exp\left[\tfrac{-n\bar{\omega}^2}{4}\right] \tag{83}$$

where we again used Hoeffding's inequality for the last line.

We bound the first term using the restated Lemma in 75:

$$\mathbb{P}\Big[|\bar{\omega}(\hat{\omega}-\omega)|\geq\frac{t}{2}\Big]\leq 2\exp\left[-\frac{n}{2}\left(\frac{t}{2\bar{\omega}}-4R_n(\mathcal{G})-\frac{2}{\sqrt{n}}\right)^2\right] \tag{84}$$

Now we let $\tilde{\epsilon}=\frac{t}{\bar{\omega}^2/2}$ to get

$$\mathbb{P}\Big[\left|\frac{\hat{\omega}}{\hat{\bar{\omega}}}-\frac{\omega}{\bar{\omega}}\right|\geq\tilde{\epsilon}\Big] \tag{85}$$

$$\leq 2\exp\left[-\frac{n}{2}\left(\frac{\tilde{\epsilon}\bar{\omega}}{4}-4R_n(\mathcal{G})-\frac{2}{\sqrt{n}}\right)^2\right]+\exp\left[\frac{-n\tilde{\epsilon}^2\bar{\omega}^4}{16\omega^2}\right]+\exp\left[\frac{-n\bar{\omega}^2}{4}\right]$$

Now we turn to the general case. Recalling that we define $\beta = \max(|\beta_1, \beta_0|)$, we have

$$\mathbb{P}\left[\left|\frac{\beta_1\hat{\omega}_1}{\hat{\bar{\omega}}_1}-\frac{\beta_0\hat{\omega}_0}{\hat{\bar{\omega}}_0}-\left(\frac{\beta_1\omega_1}{\bar{\omega}_1}-\frac{\beta_0\omega_0}{\bar{\omega}_0}\right)\right|\geq\epsilon\right]\leq$$

$$\mathbb{P}\left[|\beta_1|\left|\frac{\hat{\omega}_1}{\hat{\bar{\omega}}_1}-\frac{\omega_1}{\bar{\omega}_1}\right|+|\beta_0|\left|\frac{\hat{\omega}_0}{\hat{\bar{\omega}}_0}-\frac{\omega_0}{\bar{\omega}_0}\right|\geq\epsilon\right]\leq$$

$$\mathbb{P}\left[\left|\frac{\hat{\omega}_1}{\hat{\bar{\omega}}_1}-\frac{\omega_1}{\bar{\omega}_1}\right|\geq\frac{\epsilon}{2\beta}\right]+\mathbb{P}\left[\left|\frac{\hat{\omega}_0}{\hat{\bar{\omega}}_0}-\frac{\omega_0}{\bar{\omega}_0}\right|\geq\frac{\epsilon}{2\beta}\right]\leq$$

$$2\exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_1}{8\beta}-4R_n(\mathcal{G})-\frac{2}{\sqrt{n}}\right)^2\right]+\exp\left[\frac{-n\epsilon^2\bar{\omega}_1^4}{64\beta\omega_1^2}\right]+\exp\left[\frac{-n\bar{\omega}_1^2}{4}\right]$$

$$+2\exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_0}{8\beta}-4R_n(\mathcal{G})-\frac{2}{\sqrt{n}}\right)^2\right]+\exp\left[\frac{-n\epsilon^2\bar{\omega}_0^4}{64\beta\omega_0^2}\right]+\exp\left[\frac{-n\bar{\omega}_0^2}{4}\right]$$

$$\leq 4\exp\left[-\frac{n}{2}\left(\frac{\epsilon\bar{\omega}_\wedge}{8\beta}-4R_n(\mathcal{G})-\frac{2}{\sqrt{n}}\right)^2\right]+2\exp\left[\frac{-n\epsilon^2\bar{\omega}_\wedge^4}{64\beta^2\omega_\vee^2}\right]$$

$$+2\exp\left[\frac{-n\bar{\omega}_\wedge^2}{4}\right]$$

where the first inequality holds by triangle inequality, the second inequality holds by the union bound, the third inequality applies 85 for $\tilde{\epsilon} = \frac{\epsilon}{2\beta}$, and the final inequality simplifies the bound using the notation $\omega_\vee = \max(\omega_1, \omega_0)$ and $\bar{\omega}_\wedge = \min(\bar{\omega}_1, \bar{\omega}_0)$. □

C.3 Auxiliary Lemmas for the Proof of Theorem 3

C.3.1 Iteration Complexity for Algorithm 4

Lemma 10. *Defining*

$$\rho := \max_{h\in\mathcal{H},\xi\in[0,B_\xi]}\max\{\widehat{\text{disparity}}(h)-\xi, -\widehat{\text{disparity}}(h)-\xi, \widehat{\text{cost}}(h)-\hat{\epsilon}\},$$

Algorithm 4 satisfies the inequality

$$\nu_t \leq \frac{B_\lambda\log(3)}{\eta t}+\eta\rho^2 B.$$

For $\eta = \frac{\nu}{2\rho^2 B_\lambda}$, Algorithm 4 will return a ν-approximate saddle point of L in at most $\frac{4\rho^2 B_\lambda^2\log(3)}{\nu^2}$ iterations. Setting $B_\xi = 1$, we observe $\rho \leq 1$, and so the iteration complexity of Algorithm 4 is $\frac{4B_\lambda^2\log(3)}{\nu^2}$.

Proof. Follows immediately from the proof of Theorem 3 in [23] and the same argument given in the proof of Lemma 3. □

C.3.2 Solution Quality for Algorithm 4

Let $\Lambda = \{\lambda \in \mathbb{R}_+^3 : \|\lambda\| \leq B_\lambda\}$. Assume we are given $\left(\hat{\xi}, \hat{Q}_h, \hat{\lambda}\right)$, which is a ν-approximate saddle point satisfying

$$L(\hat{\xi}, \hat{Q}_h, \hat{\lambda}) \leq L(\xi, Q_h, \hat{\lambda}) + \nu \text{ for all } Q_h \in \Delta(\mathcal{H}), \xi \in [0, B_\xi]$$
$$L(\hat{\xi}, \hat{Q}_h, \hat{\lambda}) \geq L(\hat{\xi}, \hat{Q}_h, \lambda) - \nu \text{ for all } \|\lambda\| \leq B_\lambda.$$

We extend Lemmas 4-6 to the problem of finding the absolute disparity minimizing model.

Lemma 11. $\left(\hat{\xi}, \hat{Q}_h, \hat{\lambda}\right)$ *satisfies*

$$\hat{\lambda}_+(\widehat{\text{disparity}}(\hat{Q}_h) - \hat{\xi}) + \hat{\lambda}_-(-\widehat{\text{disparity}}(\hat{Q}_h) - \hat{\xi}) + \hat{\lambda}_{cost}\left(\widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon}\right)$$
$$\geq B_\lambda \max\{\widehat{\text{disparity}}(\hat{Q}_h) - \hat{\xi}, -\widehat{\text{disparity}}(\hat{Q}_h) - \hat{\xi}, \widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon}\} - \nu.$$

Proof. The argument is the same as the proof of Lemma 4. □

Lemma 12. *The value $\hat{\xi}$ satisfies*

$$\hat{\xi} \leq \xi + 2\nu$$

for any ξ such that there exists Q_h satisfying $\widehat{\text{disparity}}(Q_h) - \xi \leq 0$, $-\widehat{\text{disparity}}(Q_h) - \xi \leq 0$ and $\widehat{\text{cost}}(Q_h) \leq \hat{\epsilon}$.

Proof. Assume the pair (ξ, Q_h) satisfies $\widehat{\text{disparity}}(Q_h) - \xi \leq 0$, $-\widehat{\text{disparity}}(Q_h) - \xi \leq 0$ and $\widehat{\text{cost}}(Q_h) \leq \hat{\epsilon}$. Since $\hat{\lambda} \geq 0$, we have that $L(\xi, Q, \hat{\lambda}) \leq \xi$. Moreover, the ν-approximate optimality conditions imply that $L(\hat{\xi}, \hat{Q}, \hat{\lambda}) \leq L(\xi, Q, \hat{\lambda}) + \nu$. Together, these inequalities imply that

$$L(\hat{\xi}, \hat{Q}, \hat{\lambda}) \leq \xi + \nu.$$

Next, we can use Lemma 11 to construct a lower bound for $L(\hat{\xi}, \hat{Q}, \hat{\lambda})$. To do so, observe that

$$\begin{aligned}
&L(\hat{\xi}, \hat{Q}, \hat{\lambda}) \\
&\geq \hat{\xi} + B_\lambda \max\{\widehat{\text{disparity}}(\hat{Q}) - \hat{\xi}, -\widehat{\text{disparity}}(\hat{Q}) - \hat{\xi}, \widehat{\text{cost}}(\hat{Q}) - \hat{\epsilon}\} - \nu \\
&\geq \hat{\xi} - \nu.
\end{aligned}$$

By combining the inequalities, $L(\hat{\xi}, \hat{Q}, \hat{\lambda}) \geq \hat{\xi} - \nu$ and $L(\hat{\xi}, \hat{Q}, \hat{\lambda}) \leq \xi + \nu$, we arrive at the claim. □

Lemma 13. *Assume the empirical cost constraint $\widehat{\text{cost}}Q_h \leq \hat{\epsilon}$ and the slack variable constraints $\widehat{\text{disparity}}(Q_h) - \xi \leq 0$ and $-\widehat{\text{disparity}}(Q_h) - \xi \leq 0$ are feasible. Then, the pair $(\hat{\xi}, \hat{Q}_h)$ satisfies*

$$\begin{aligned}
\widehat{\text{disparity}}(\hat{Q}_h) - \hat{\xi} &\leq \frac{B_\xi + 2\nu}{B_\lambda}, \\
-\widehat{\text{disparity}}(\hat{Q}_h) - \hat{\xi} &\leq \frac{B_\xi + 2\nu}{B_\lambda}.
\end{aligned}$$

Proof. Let ξ be a feasible value of the slack variable such that there exists Q_h satisfying $\widehat{\text{cost}}(Q_h) \leq \hat{\epsilon}$ and the slack variable constraints $\widehat{\text{disparity}}(Q_h) - \xi \leq 0$, $-\widehat{\text{disparity}}(Q_h) - \xi \leq 0$. Recall from the Proof of Lemma 12, we showed that

$$\begin{aligned}
&\hat{\xi} + B_\lambda \max\{\widehat{\text{disparity}}(\hat{Q}_h) - \hat{\xi}, -\widehat{\text{disparity}}(\hat{Q}_h) - \hat{\xi}, \widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon}\} - \nu \\
&\leq L(\hat{\xi}, \hat{Q}_h, \hat{\lambda}) \leq \xi + \nu.
\end{aligned}$$

Therefore, it is immediate that

$$B_\lambda \max\{\widehat{\text{disparity}}(\hat{Q}_h) - \hat{\xi}, -\widehat{\text{disparity}}(\hat{Q}_h) - \hat{\xi}, \widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon}\} \leq \left(\xi - \hat{\xi}\right) + 2\nu,$$

and so

$$\begin{aligned}
B_\lambda \left(\widehat{\text{disparity}}(\hat{Q}_h) - \hat{\xi}\right) &\leq \left(\xi - \hat{\xi}\right) + 2\nu, \\
B_\lambda \left(-\widehat{\text{disparity}}(\hat{Q}_h) - \hat{\xi}\right) &\leq \left(\xi - \hat{\xi}\right) + 2\nu.
\end{aligned}$$

Since $\xi \in [0, B_\xi]$, we can bound $\xi - \hat{\xi}$ by B_ξ. The result follows. □

Lemma 14. *Assume the empirical cost constraint* $\widehat{\text{cost}}(Q_h) \le \hat{\epsilon}$ *and the slack variable constraints* $\widehat{\text{disparity}}(Q_h) - \xi \le 0$, $-\widehat{\text{disparity}}(Q_h) - \xi \le 0$ *are feasible. Then the distribution* $\hat{Q}_h$ *satisfies*

$$\widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon} \le \frac{B_\xi + 2\nu}{B_\lambda}.$$

Proof. The proof is analogous to the proof of Lemma 13. □

Lemma 15. *Suppose that* (ξ, Q_h) *is a feasible solution to the empirical version of (26). Then, the solution* $(\hat{\xi}, \hat{Q}_h)$ *returned by Algorithm 4 satisfies*

$$\begin{aligned}
&\hat{\xi} \le \xi + 2\nu, \\
&\widehat{\text{disparity}}(\hat{Q}_h) - \hat{\xi} \le \frac{B_\xi + 2\nu}{B_\lambda}, \\
&-\widehat{\text{disparity}}(\hat{Q}_h) - \hat{\xi} \le \frac{B_\xi + 2\nu}{B_\lambda} \\
&\widehat{\text{cost}}(\hat{Q}_h) - \hat{\epsilon} \le \frac{B_\xi + 2\nu}{B_\lambda}.
\end{aligned}$$

Proof. The proof follows from Lemmas 13-14. If the algorithm returns $null$, then these inequalities are vacuously satisfied. □

D Additional Details on the Consumer Lending Data

D.1 Construction of IRSD for SA4 Regions

As discussed in § 6, we focus our analysis on predictive disparities across SA4 geographic regions within Australia. We use the Australian Bureau of Statistics' Index of Relative Socioeconomic Disadvantage (IRSD) to define socioeconomically disadvantaged SA4 regions. In this section, we provide additional details on the IRSD and our construction of the population-weighted IRSD for SA4 geographic regions.

The IRSD is calculated for SA2 regions, which are more granular statistical areas used by the ABS, by aggregating sixteen variables that were collected in the 2016 Australian census. These variables include, for example, the fraction of households making less than $26,000 AUD, the fraction of households with no internet access, and the fraction of residents who do not speak English well. Higher scores on the IRSD are associated with less socioeconomically disadvantaged regions, and conversely, lower scores on the IRSD are associated with more socioeconomically disadvantaged regions. The full list of variables that are included in the IRSD and complete details on how the IRSD is constructed is provided in [48].

Because the IRSD is constructed for SA2 regions, we first aggregate this index to SA4 regions. We construct an aggregated IRSD for each SA4 region by constructing a population-weighted average of the IRSD for all SA2 regions that fall within each SA4 region. This delivers a quantitative measure of which SA4 regions are the most and least socioeconomically disadvantaged. For example, the bottom ventile (i.e., the 20th ventile) of SA4 regions based upon the population-weighted average IRSD (i.e., the least socioeconomically disadvantaged SA4 regions) are regions associated with Sydney and Perth. The top ventile (i.e., the 1st ventile) of SA4 regions based upon the population-weighted average IRSD (i.e., the most socioeconomically disadvantaged SA4 regions) are regions associated with the Australian outback such as the Northern territory outback and the Southern Australia outback. Figure 4 provides a map of SA4 regions in Australia, in which colors SA4 regions classified as socioeconomically disadvantaged in blue.

E Additional Experiments

In this section, we present additional results for our consumer lending experiment, recidivism risk prediction experiments and apply our methods to analyze the UCI adult income dataset [50].

E.1 Consumer Lending Risk Scores: Additional Results

Figure 5 shows the mean square error (MSE) against predictive disparity for the KGB, RIE, IE benchmarks and disparity-minimizing variants on held-out test data. The qualitative patterns are the same as Figure 1 in § 6 of the

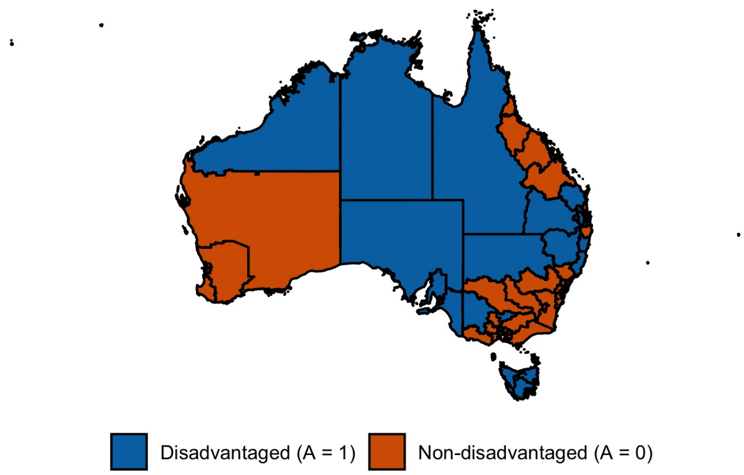


Figure 4: SA4 regions in Australia. We classify SA4 regions as being "socioeconomically disadvantaged" (red) and "non-socioeconomically disadvantaged" (blue) based on the Index of Relative Socioeconomic Disadvantage (IRSD).

main text. Evaluation on all applicants shows that the RIE and IE methods outperform and achieve lower disparity rates than the KGB model trained only on funded data. We again observe that evaluation on only funded applications is misleading as it suggests that the KGB models have comparable MSE and it drastically underestimates predictive disparities for all models.

We notice that Figure 1 in § 6 of the main text and Figure 5 shows that the disparity-minimizing KGB model appears to produce larger predictive disparities than the benchmark KGB model. This is likely due to generalization error on the held-out test data. To verify this hypothesis, Figure 6 shows the MSE against predictive disparity for the KGB, RIE, IE benchmarks and disparity-minimizing variants on the training data. Indeed among funded applicants in the train data, the disparity-minimizing KGB models produce smaller absolute predictive disparities than the benchmark KGB model (second row). We continue to observe in the training data that evaluation on only funded applications is misleading.

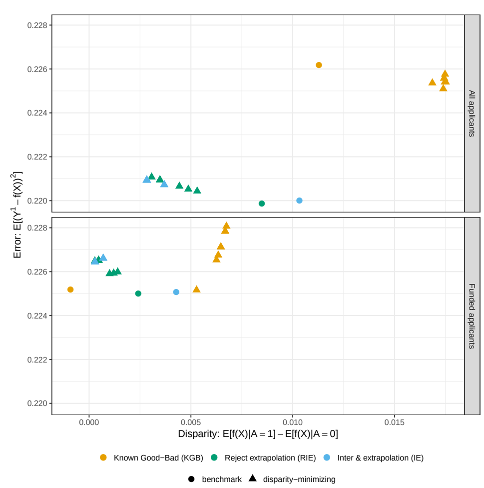


Figure 5: Mean square error (MSE) with respect to the synthetic outcome $\tilde{Y}_i$ against disparity in the average risk prediction for the disadvantaged ($A_i = 1$) vs. advantaged ($A_i = 0$) groups in held-out test data. The first row evaluates each method on all applicants and and the second row evaluates each method on funded (known) applicants only. See § 6 and § E.1 for further details.

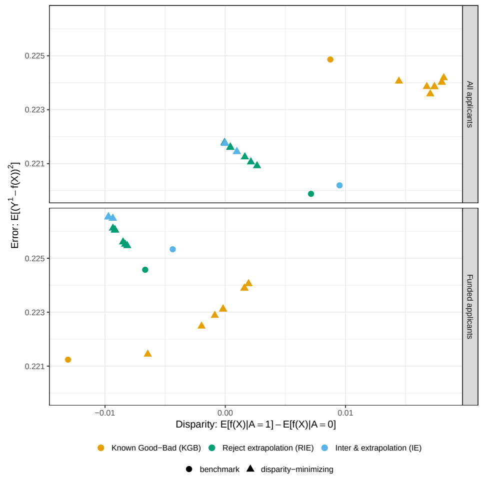
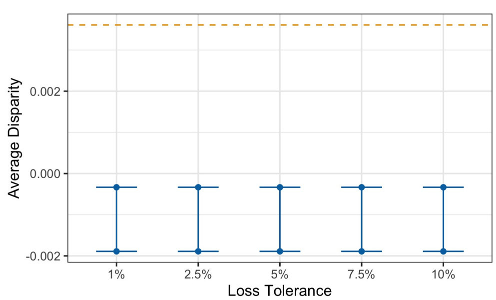


Figure 6: Mean square error (MSE) with respect to the synthetic outcome $\tilde{Y}_i$ against disparity in the average risk prediction for the disadvantaged ($A_i = 1$) vs. advantaged ($A_i = 0$) groups in the training data. The first row evaluates each method on all applicants and and the second row evaluates each method on funded (known) applicants only. See § 6 and § E.1 for further details.

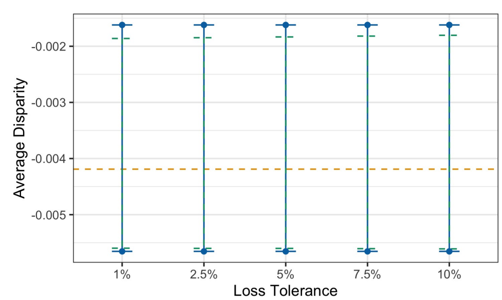
E.2 Recidivism Risk Prediction: Additional Results

E.2.1 Bounded Group Loss Disparities across Black and White Defendants

Figure 7 plots the range of relative bounded group loss disparities over the test and train set when the parameter ϵ is calibrated using the COMPAS score. Figure 7a plots the range of relative disparities over the test set: the blue error bars plot the relative disparities associated with the linear program reduction (§ 4.2) and the orange dashed line plots the relative disparity associated with the COMPAS score. Figure 7b plots the range of relative disparities over the train set: the blue error bars plot the relative disparities associated with the linear program reduction, the green error bars plot the relative disparities associated with the stochastic prediction function returned by Algorithm 5 and the orange dashed line plots the relative disparity associated with the COMPAS score.



(a) Test Data

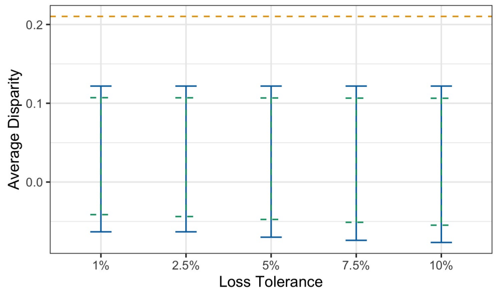


(b) Train Data

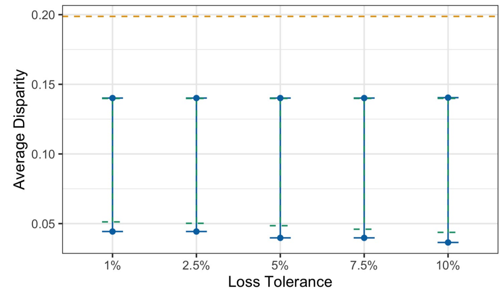
Figure 7: Range of relative bounded group loss disparities between black defendants ($A = 1$) and white defendants ($A = 0$) over the train set as the loss tolerance varies. The loss tolerance is calibrated using the COMPAS score as the benchmark model. See § 7 and § E.2.1 for further details.

E.2.2 Train Set Performance

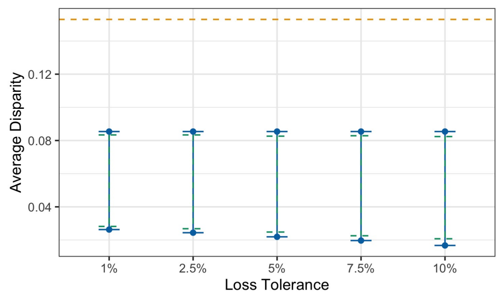
Figure 8 plots the range of predictive disparities over the train set when the parameter ϵ is calibrated using the COMPAS score. Notice that in the train dataset, the range of disparities produced by the linear program reduction closely track the range of disparities produced by the stochastic prediction function returned by Algorithm 1. The qualitative patterns that we observed in the test set are the same in the train set.



(a) Statistical Parity



(b) Balance for the Positive Class



(c) Balance for the Negative Class

Figure 8: Range of predictive disparities between black defendants ($A = 1$) and white defendants ($A = 0$) as the loss tolerance varies. The loss tolerance is calibrated using the COMPAS score as the benchmark model. See § 7 and § E.2.2 for further details.

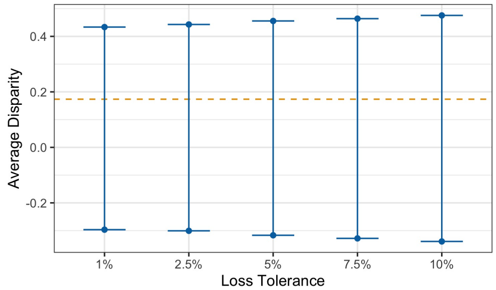
E.2.3 Results for Predictive Disparities across Young and Older Defendants

We next examined the range of predictive disparities between defendants that are younger than 25 years old ($A = 1$) and defendants older than 25 years old ($A = 0$). We focus on the range of predictive disparities that could be generated by a risk score that is constructed using logistic regression on a quadratic polynomial of the defendant's age and number of prior offenses. We again calibrate the loss tolerance parameter ϵ such that (3) constructs the fairness frontier over all models that achieve a logistic regression loss within $\{1\%, 2.5\%, 5\%, 7.5\%, 10\%\}$ of the COMPAS score. We provide the results for the statistical parity, balance for the positive class, balance for the negative class and bounded group loss disparity measure and evaluate the models over the test set. Figure 9 plots the range of predictive disparities over the test set when the parameter ϵ is calibrated using the COMPAS score: the blue error bars plot the relative disparities associated with the linear program reduction (§ 4.2) and the orange dashed line plots the relative disparity associated with the COMPAS score.

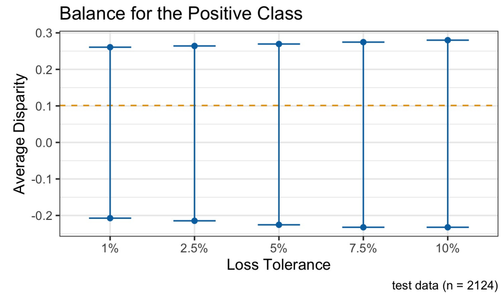
Figure 10 plots the range of predictive disparities over the train set: the blue error bars plot the relative disparities associated with the linear program reduction (Section 4.2), the green error bars plot the relative disparities associated with the stochastic prediction function returned by Algorithm 5 and the orange dashed line plots the relative disparity associated with the COMPAS score.

E.3 UCI Adult Income Dataset

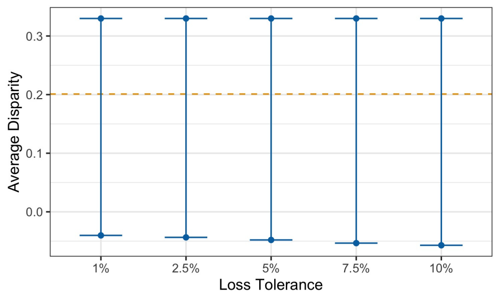
We use the UCI adult income dataset [50] and predict whether an individual's annual income is more than \$50,000 (Y). We examine the range of predictive disparities across females $(A = 1)$ and males $(A = 0)$ that may be generated by a prediction function constructed using logistic regression with features that one-hot encode the education level, occupation, race and marital status of the individual. We calibrate the parameter ϵ by setting it such that (3) constructs the range of predictive disparities over all models that achieve a logistic regression loss within {1%, 2.5%, 5%, 7.5%, 10%} of the loss-minimizing prediction function. In this analysis, we created a smaller version of the UCI adult income dataset by randomly subsampling 2500 observations for the train set and 2500 observations for the test set.



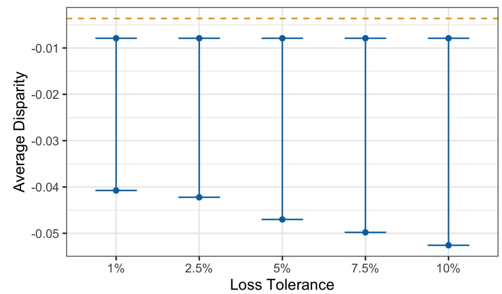
(a) Statistical Parity



(b) Balance for the Positive Class

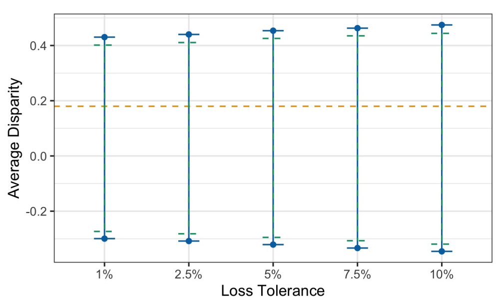


(c) Balance for the Negative Class

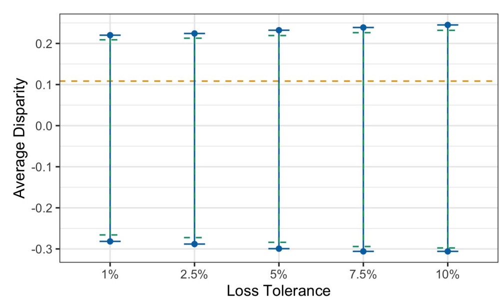


(d) Bounded Group Loss

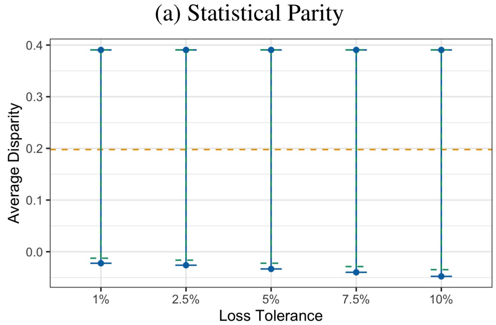
Figure 9: Range of relative disparities over the test set between young defendants ($A = 1$) and older defendants ($A = 0$) in the ProPublica COMPAS dataset as the loss tolerance varies. The loss tolerance is calibrated using the COMPAS score as the benchmark model. See § E.2.3 of the Supplement for further details.



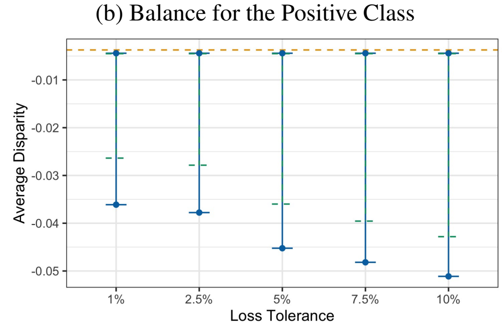
(a) Statistical Parity



(b) Balance for the Positive Class

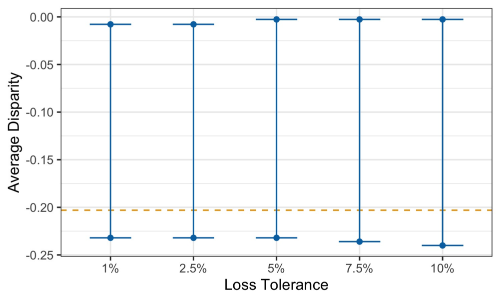


(c) Balance for the Negative Class

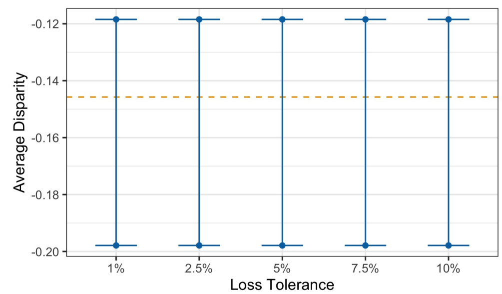


(d) Bounded Group Loss

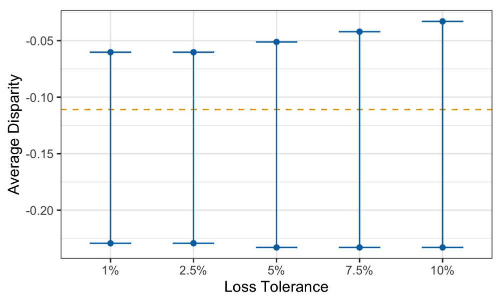
Figure 10: Range of relative disparities over the train set between young defendants ($A = 1$) and older defendants ($A = 0$) in the ProPublica COMPAS as the loss tolerance varies. The loss tolerance is calibrated using the COMPAS score as the benchmark model. See § E.2.3 of the Supplement for further details.



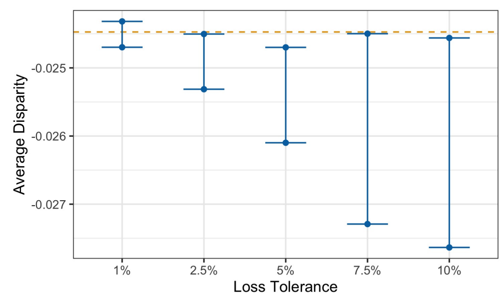
(a) Statistical Parity



(b) Balance for the Positive Class



(c) Balance for the Negative Class

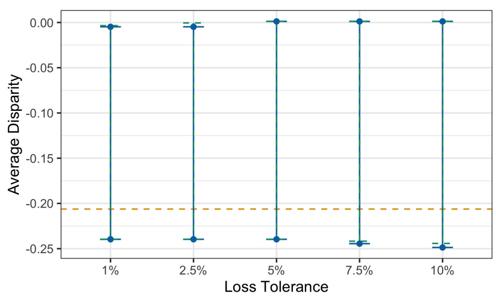


(d) Bounded Group Loss

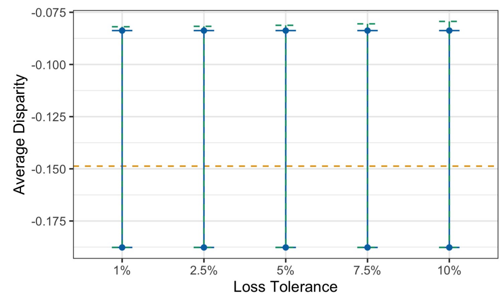
Figure 11: Range of relative disparities over the test set between females ($A = 1$) and males ($A = 0$) in predicting whether an individual's annual income is more than \$50,000 in the UCI adult income dataset as the loss tolerance varies. The loss tolerance is calibrated using the loss minimizing model as the benchmark. See § E.3 of the Supplement for further details.

E.3.1 Results

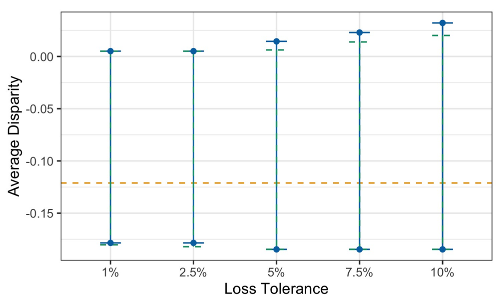
Figure 11 plots the range of predictive disparities over the test set for the UCI adult income dataset. The range of relative disparities associated with the linear program reduction of (Section 4.2) are constructed over the test set and are plotted in the blue error bars. The relative disparity produced the logistic regression loss minimizing model in the test set is plotted in the orange dashed line. We observe that there is a large range of predictive disparities that are possible over the set of good models in the UCI adult income dataset. Figure 12 plots the range of predictive disparities over the train set for the UCI adult income dataset. The solid blue error bars plots the range of relative disparities associated the linear program reduction (§ 4.2). The green dashed line plots the range of relative disparities associated with the stochastic prediction function produced by Algorithm 1 and Algorithm 5. The orange dashed line plots the relative disparity associated with the logistic regression loss minimizing model in the train set. For statistical parity, balance for the positive class and balance for the negative class, the disparities reported by the stochastic prediction function and linear program reduction track each other quite closely. In contrast, there is a large gap between the disparities reported by the stochastic prediction function and linear program reduction for bounded group loss. This behavior appears to be driven by the initial steps of Algorithm 5, which returns prediction functions that perform quite poorly in terms of average loss (i.e., average loss larger than the loss tolerance but still feasible after accounting for the algorithm's slackness). These initial poorly performing models are extreme outliers and have a strong effect on the performance of the stochastic prediction. In contrast, the linear programming reduction only selects among the prediction functions that satisfy the loss tolerance constraint. This generates sizeable differences between the overall loss of the linear program reduction and the stochastic prediction function.



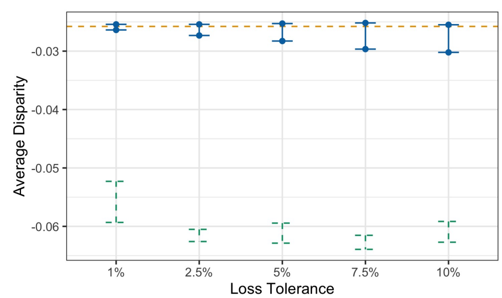
(a) Statistical Parity



(b) Balance for the Positive Class



(c) Balance for the Negative Class



(d) Bounded Group Loss

Figure 12: Range of relative disparities between females ($A = 1$) and males ($A = 0$) over the train set in predicting whether an individual's annual income is more than \$50,000 in the UCI adult income dataset as the loss tolerance varies. The loss tolerance is calibrated using the loss minimizing model as the benchmark model. See § E.3 of the Supplement for further details.