Fair Regression: Quantitative Definitions and Reduction-based Algorithms

Alekh Agarwal [1] **Miroslav Dudík** [2] **Zhiwei Steven Wu** [3]

Abstract

In this paper, we study the prediction of a real-valued target, such as a risk score or recidivism rate, while guaranteeing a quantitative notion of fairness with respect to a protected attribute such as gender or race. We call this class of problems *fair regression*. We propose general schemes for fair regression under two notions of fairness: (1) statistical parity, which asks that the prediction be statistically independent of the protected attribute, and (2) bounded group loss, which asks that the prediction error restricted to any protected group remain below some pre-determined level. While we only study these two notions of fairness, our schemes are applicable to arbitrary Lipschitz-continuous losses, and so they encompass least-squares regression, logistic regression, quantile regression, and many other tasks. Our schemes only require access to standard risk minimization algorithms (such as standard classification or least-squares regression) while providing theoretical guarantees on the optimality and fairness of the obtained solutions. In addition to analyzing theoretical properties of our schemes, we empirically demonstrate their ability to uncover fairness–accuracy frontiers on several standard datasets.

1. Introduction

As machine learning touches increasingly critical aspects of our life, including education, healthcare, criminal justice and lending, there is a growing focus to ensure that the algorithms treat various subpopulations fairly (see, e.g., Barocas & Selbst, 2016; Podesta et al., 2014; Corbett-Davies & Goel, 2018; and references therein). These questions have been particularly extensively researched in the context of classification, where several quantitative measures of fairness have been proposed (Berk et al., 2017; Chouldechova, 2017; Hardt et al., 2016; Kleinberg et al., 2017), leading to a variety of algorithms that aim to satisfy them (see, e.g., Corbett-Davies & Goel, 2018, for an overview of the literature).

These classifier-based formulations appear to fit the settings where the decision space is discrete and small, such as accept/reject decisions in hiring, school admissions, or lending. However, in practice, the decision makers work with tools that estimate a continuous quantity, such as success on the job, GPA in the first year of college, or risk of default on a loan. Predictions of these quantities are treated as scores, which are used by human decision makers, perhaps in the context of a partly automated workflow, to reach final decisions (see, e.g., Waters & Miikkulainen, 2014; US Federal Reserve, 2007; Northpointe, 2010; Lowenkamp et al., 2012). While, in principle, a fair classification tool could be used to recommend the yes/no decision directly, such tools are often resisted by practitioners, because they limit their autonomy, whereas ranking or scoring tools do not have this drawback (Veale et al., 2018). In such situations, it is desirable to work with real-valued scores that satisfy some notion of fairness. Yet, despite ample motivation and use cases, the prior work on designing fair continuous predictors is quite limited in its scope compared with the generality of methods for fair classification (e.g., Hardt et al., 2016; Agarwal et al., 2018).

This paper seeks to diminish this gap by developing efficient algorithms for a substantially broader set of regression tasks and model classes than done before, in many cases providing the first method with theoretical performance guarantees.

We consider the problem of predicting a real-valued target, where the prediction quality is measured by any Lipschitz-continuous loss function. Each example contains a *protected attribute*, such as race or gender, with respect to which we seek to guarantee fairness. We study two definitions of fairness from previous literature: *statistical parity* (SP), which asks that the prediction be statistically independent of the protected attribute, and *bounded group loss* (BGL), which asks that the prediction error restricted to any protected group stay below some pre-determined level. We define fair regression as the task of minimizing the expected loss of our real-valued predictions, subject to either of these fairness constraints. By choosing the appropriate loss, we

[1]Microsoft Research, Redmond, WA [2]Microsoft Research, New York, NY [3]University of Minnesota, Minneapolis, MN. Correspondence to: A. Agarwal <alekha@microsoft.com>, M. Dudík <mdudik@microsoft.com>, Z. S. Wu <zsw@umn.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

obtain a wide range of standard prediction tasks including least-squares, logistic, Poisson, and quantile regression (with labels and predictions restricted to a bounded set to obtain Lipschitz continuity). While we seek to solve the regression tasks under fairness constraints, our schemes only require access to standard risk minimization algorithms such as standard classification or least-squares regression.

Several prior works also seek predictors that exhibit some form of independence from the protected attribute similar to statistical parity. Calders et al. (2013), Johnson et al. (2016) and Komiyama et al. (2018) consider a more limited form of independence, expressed via a small number of moment constraints, such as lack of correlation, and design specific algorithms for linear least squares. Berk et al. (2017) study notions of individual and group fairness specialized to linear regression. Pérez-Suay et al. (2017) seek zero correlation in a reproducing kernel Hilbert space (RKHS), which can capture statistical independence, but it only yields predictors in the same RKHS and the loss is limited to least squares. Kamishima et al. (2012) and Fukuchi et al. (2013) seek to fit a probabilistic model that satisfies statistical independence, but they do not present efficient algorithms or statistical guarantees. In contrast, we consider full statistical independence, arbitrary model classes and Lipschitz losses, and our algorithms are efficient and come with statistical guarantees.

Our second fairness definition, bounded group loss, fits into the general framework of Alabi et al. (2018), whose goal is to minimize a general function of group-wise prediction losses, but their algorithm is less efficient (albeit still polynomial), and they do not provide statistical guarantees.

We design a separate algorithm for each of the two fairness definitions. For BGL, our insight is that the problem of loss minimization subject to a loss bound in each subpopulation can be algorithmically reduced to a weighted loss minimization problem for which standard approaches exist. For SP, the main obstacle is that the number of constraints is uncountable. Here, the main insight that allows us to design and analyze the algorithm is that if we discretize the real-valued prediction space, then the task of fair regression can be reduced to cost-sensitive classification under certain constraints. We build on the recent work of Agarwal et al. (2018), and use the special structure of our discretization scheme to develop several algorithms reducing to standard classification or regression problems without fairness constraints. We provide theoretical results to bound the computational cost, generalization error and fairness violation of the returned predictor for both of our fairness measures with arbitrary Lipschitz-continuous loss functions and with arbitrary regression-function classes of bounded complexity, again building on the analysis of Agarwal et al. (2018). Prior works in the regression setting lack such guarantees.

Empirically, we evaluate our method on several standard datasets, on the tasks of least-squares and logistic regression under statistical parity, with linear and tree-ensemble learners, and compare it with the unconstrained baselines as well as the technique of Johnson et al. (2016). Our method uncovers fairness–accuracy frontiers and provides the first systematic scheme for enforcing fairness in a significantly broader class of learning problems than prior work.

Usage guidelines. We envision the use of our algorithms in uncovering fairness–accuracy frontiers in a variety of applications. Any substantial tradeoffs along the frontier need to be analyzed. They might point to data issues requiring non-algorithmic interventions, such as gathering of additional (less biased) data or introduction of new features (Chen et al., 2018). As with other algorithmic fairness tools, in order to successfully use our algorithms in practice, it is essential to consider the societal context of the application (Selbst et al., 2018). In some contexts, the best fairness intervention might be to avoid a technological intervention altogether.

2. Problem Formulation

We consider a general prediction setting where the training examples consist of triples (X, A, Y), where $X \in \mathcal{X}$ is a feature vector, $A \in \mathcal{A}$ is a protected attribute and $Y \in \mathcal{Y} \subseteq [0, 1]$ is the label. Throughout, we focus on the protected attribute taking a small number of discrete values, i.e., $\mathcal{A}$ is finite, but $\mathcal{X}$ is allowed to be continuous and high-dimensional. We make no specific assumptions about whether the protected attribute is included in the feature vector X or not; also the set of labels $\mathcal{Y}$ can be discrete (but embedded in $[0, 1]$) or continuous. Given a set of predictors $\mathcal{F}$ containing functions $f : \mathcal{X} \to [0, 1]$, our goal is to find $f \in \mathcal{F}$ which is accurate in predicting Y given X while satisfying some fairness condition such as statistical parity or bounded group loss (formally defined below). Note that the functions f do not explicitly depend on A unless it is included in X.

The main departure from prior works on classification is that Y as well as $f(X)$ are allowed to be real-valued rather than just categorical. The accuracy of a prediction $f(X)$ on a label Y is measured by the loss $\ell(Y, f(X))$. The loss function $\ell : \mathcal{Y} \times [0, 1] \to [0, 1]$ is required to be 1-Lipschitz under the ℓ_1 norm,[1] that is:

$$|\ell(y, u) - \ell(y', u')| \leq |y - y'| + |u - u'| \quad \text{for all } y, y', u, u'.$$

Example 1 (Least-squares regression)**.** *The prediction of GPA in the first year of college can be cast as a regression problem where the label y is the normalized GPA so that $\mathcal{Y} = [0, 1]$, and the error is measured by the square loss $\ell(y, f(x)) = (y - f(x))^2/2$. Since $y, f(x) \in [0, 1]$, the loss is bounded and 1-Lipschitz.*

[1] Our algorithms primarily use covers for y and u such that $\ell(y, u)$ can be approximated using corresponding elements from the cover. We skip this generalization to keep presentation simple.

Example 2 (Logistic regression)**.** *Consider a system for screening job applicants based on the likelihood of an offer upon interview. We train this system using past data of interviewed candidates where X describes their features and $Y \in \{0,1\}$ the hiring decision. The scoring function f can be chosen to maximize the likelihood for the logistic model $p_f(Y = 1 \mid x) = 1/(1 + e^{-f(x)})$. Since we require that $f(x) \in [0,1]$, in order to approximate the full range of probabilities, we use a scaled and shifted version $p_f(Y = 1 \mid x) = 1/\big(1 + e^{-C(2f(x)-1)}\big)$ with some $C > 1$, giving probabilities in the range $[1/(1+e^C),\ 1/(1+e^{-C})]$. The loss is a rescaled version of the negative log likelihood to ensure the boundedness and 1-Lipschitz conditions: $\ell(y, f(x)) = \log\big(1+e^{-C(2y-1)(2f(x)-1)}\big)/\big(2\log(1+e^C)\big)$. Here the label is binary, but the prediction is real-valued.*

2.1. Fairness Definitions

We consider two quantitative definitions of fairness appearing in prior work on fair classification and regression.

The first definition, called *statistical* (or demographic) *parity*, says that the prediction should be independent of the protected attribute. In classification, it corresponds to the practice of affirmative action (see, e.g., Holzer & Neumark, 2006, and references therein) and it is also invoked to address disparate impact under the US Equal Employment Opportunity Commission's "four-fifths rule," which requires that the "selection rate for any race, sex, or ethnic group [must be at least] four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate."[2]

Definition 1 (Statistical parity—SP)**.** *A predictor f satisfies statistical parity under a distribution over (X, A, Y) if $f(X)$ is independent of the protected attribute A. Since $f(X) \in [0,1]$, this is equivalent to $\mathbb{P}[f(X) \geq z \mid A = a] = \mathbb{P}[f(X) \geq z]$ for all $a \in \mathcal{A}$ and $z \in [0,1]$.*[3]

The characterization through the properties of the CDF of $f(X)$ is particularly useful when $f(X)$ can take any real values in $[0,1]$, because it allows us to design efficient algorithms. It also makes it obvious that if f satisfies SP, then any classifier induced by thresholding f will also satisfy SP.

Our second fairness definition, called *bounded group loss*, formalizes the requirement that the predictor's loss remain below some acceptable level for each protected group. In settings such as speech or face recognition, this corresponds to the requirement that all groups receive good service (cf. Buolamwini & Gebru, 2018). In other settings, such as lending and hiring, it aims to prevent situations when the predictor has a high error on some of the groups (cf. Section 3.3 of Corbett-Davies & Goel, 2018).

Definition 2 (Bounded group loss—BGL)**.** *A predictor f satisfies bounded group loss at level ζ under a distribution over (X, A, Y) if $\mathbb{E}[\ell(Y, f(X)) \mid A = a] \leq \zeta$ for all $a \in \mathcal{A}$.*

Hence, fair regression with BGL minimizes the overall loss, while controlling the worst loss on any protected group. By Lagrangian duality, this is equivalent to minimizing the worst loss on any group while maintaining good overall loss (referred to as max-min fairness). Unlike *overall accuracy equality* in classification (Dieterich et al., 2016), which requires the losses on all groups to be equal, BGL does not force an artificial decrease in performance on every group just to match the hardest-to-predict group. BGL can be used as a diagnostic for the potential shortcomings of a chosen featurization or dataset. If it is not possible to achieve a loss below ζ on some group, then to achieve fairness we need to collect more data for that group, or develop more informative features for individuals in that group.

2.2. Fair Regression

We begin by defining the problem of fair regression as the minimization of the expected loss $\mathbb{E}[\ell(Y, f(X))]$ over $f \in \mathcal{F}$, while guaranteeing SP or BGL. However, to achieve better fairness–accuracy tradeoffs we then generalize this to the case of randomized predictors.

Statistical parity. Similar to prior works on fair classification (Agarwal et al., 2018), it is frequently desirable to have a tunable knob for navigating the fairness-accuracy tradeoff, such as ζ in the definition of bounded group loss. To allow such a tradeoff in SP, we consider slack parameters ε_a for each attribute and define the fair regression task under SP as

$$\min_{f \in \mathcal{F}} \mathbb{E}\left[\ell(Y, f(X))\right] \quad \text{such that } \forall a \in \mathcal{A}, z \in [0,1]:$$
$$\Big|\mathbb{P}[f(X) \geq z \mid A = a] - \mathbb{P}[f(X) \geq z]\Big| \leq \varepsilon_a. \tag{1}$$

The slack ε_a bounds the allowed departure of the CDF of $f(X)$ conditional on $A = a$ from the CDF of $f(X)$. The difference between CDFs is measured in the ℓ_∞ norm corresponding to the Kolmogorov-Smirnov statistic (Lehmann & Romano, 2006). Choosing different ε_a allows us to vary the strength of constraint across different protected groups.

Bounded group loss. In this case, the constrained optimization formulation follows directly from the definition. For the sake of flexibility, we allow specifying a different bound ζ_a for each attribute value, leading to the formulation

$$\min_{f \in \mathcal{F}} \mathbb{E}\big[\ell(Y, f(X))\big]$$
$$\text{such that } \forall a \in \mathcal{A}: \quad \mathbb{E}\big[\ell(Y, f(X)) \,\big|\, A = a\big] \leq \zeta_a. \tag{2}$$

[2] See the Uniform Guidelines on Employment Selection Procedures, 29 C.F.R. §1607.4(D) (2015).

[3] A standard definition of statistical independence requires that $\mathbb{P}[f(X) \in S \mid A = a] = \mathbb{P}[f(X) \in S]$ for all measurable sets S. Since $f(X)$ is a real-valued random variable under Borel σ-algebra, it is fully characterized by its cumulative distribution function, and so it suffices to consider sets $S = [0, z]$ for $z \in [0,1]$ (see, e.g., Theorem 10.49 of Aliprantis & Border, 2006).

Randomized predictors. Similar to fair classification, in order to achieve better fairness–accuracy tradeoffs, we consider randomized predictors which first pick f according to some distribution Q and then predict according to f. We first introduce additional notation for the objective and constraints appearing in (1) and (2):

$$\begin{aligned}
&\mathrm{loss}(f) \coloneqq \mathbb{E}[\ell(Y, f(X))],\\
&\gamma_{a,z}(f) \coloneqq \mathbb{P}[f(X) \geq z \mid A = a] - \mathbb{P}[f(X) \geq z].\\
&\gamma_a^{\mathrm{BGL}}(f) \coloneqq \mathbb{E}\big[\ell(Y, f(X)) \,\big|\, A = a\big].
\end{aligned}$$

For a randomized predictor represented by a distribution Q, we have $\mathrm{loss}(Q) = \sum_f Q(f)\mathrm{loss}(f)$, $\gamma_{a,z}(Q) = \sum_f Q(f)\gamma_{a,z}(f)$, and $\gamma_a^{\mathrm{BGL}}(Q) = \sum_f Q(f)\gamma_a^{\mathrm{BGL}}(f)$.

Thus, for SP we seek to solve

$$\min_{Q \in \Delta(\mathcal{F})} \mathrm{loss}(Q) \text{ s.t. } \big|\gamma_{a,z}(Q)\big| \leq \varepsilon_a \;\; \forall a \in \mathcal{A}, z \in [0,1], \quad (3)$$

where $\Delta(\mathcal{F})$ is the set of all probability distributions over $\mathcal{F}$. For bounded group loss, we similarly seek to solve

$$\min_{Q \in \Delta(\mathcal{F})} \mathrm{loss}(Q) \text{ s.t. } \gamma_a^{\mathrm{BGL}}(Q) \leq \zeta_a \;\; \forall a \in \mathcal{A}. \quad (4)$$

3. Supervised Learning Oracles

In this paper, we show how to transform the fair regression problem into three standard learning problems: cost-sensitive classification, weighted least-squares regression, or weighted risk minimization under ℓ (without fairness constraints). All of these learning problems allow different costs per example, which helps incorporate fairness. The specific algorithms to solve these tasks are termed *supervised learning oracles*. These oracles are typically available for representations where regression or classification *without fairness constraints* can be solved, and we show some typical examples in our empirical evaluation.

(1) Risk minimization under ℓ. This is the most natural oracle as it implements loss minimization without fairness constraints. Given a dataset $\{(W_i, X_i, Y_i)\}_{i=1}^n$ where W_i are non-negative weights, the oracle returns $f \in \mathcal{F}$ that minimizes the weighted empirical risk: $\sum_{i=1}^n W_i \ell(Y_i, f(X_i))$.

(2) Square loss minimization. Even when the accuracy is measured by ℓ, we typically have access to a weighted least-squares learner for the same class $\mathcal{F}$. This oracle takes the data $\{(W_i, X_i, Y_i)\}_{i=1}^n$ and returns $f \in \mathcal{F}$ that minimizes the weighted squared loss: $\sum_{i=1}^n W_i\big(Y_i - f(X_i)\big)^2$.

(3) Cost-sensitive classification (CS). Our third type of oracle optimizes over classifiers $h : \mathcal{X}' \to \{0,1\}$ from some class $\mathcal{H}$. As input, we are given a dataset $\{(X_i', C_i)\}_{i=1}^n$, where X_i' is a feature vector and C_i indicates the difference between the cost (i.e., the loss) of predicting 1 versus 0; positive C_i means that 0 is favored, negative C_i means that 1 is favored. The goal is to find a classifier $h \in \mathcal{H}$, which minimizes the empirical cost relative to the cost of predicting all zeros: $\sum_{i=1}^n C_i h(X_i')$.

CS reduces to *weighted binary classification* on the data $\{(W_i, X_i', Y_i)\}_{i=1}^n$ with $Y_i = \mathbf{1}\{C_i \leq 0\}$ and $W_i = |C_i|$, where we minimize $\sum_{i=1}^n W_i \mathbf{1}\{h(X_i') \neq Y_i\}$. Weighted classification oracles exist for many classifier families $\mathcal{H}$.

In this paper we consider classifiers obtained by thresholding regressors $f \in \mathcal{F}$. We define $\mathcal{X}' = \mathcal{X} \times \mathbb{R}$ where the new feature specifies a threshold. Our classifiers act on $x' = (x, z)$ and predict $h_f(x, z) = \mathbf{1}\{f(x) \geq z\}$. This structure of classifiers naturally arises from the SP constraints. We assume access to a CS oracle for $\mathcal{H} = \{h_f : f \in \mathcal{F}\}$. While cost-sensitive learners for this representation might not be available off the shelf, learners based on optimization, such as (stochastic) gradient-based learners, can usually be adapted to this structure. In particular, it is easy to adapt learners for logistic regression, SVMs or neural nets.

4. Fair Regression under Statistical Parity

We next show how to solve the fair regression problem (3) using a CS oracle. We begin by recasting the problem (3) as a *constrained (and cost-sensitive) classification problem*, which we then solve via the reduction approach of Agarwal et al. (2018), by repeatedly invoking the CS oracle.

We proceed in two steps. First we discretize our prediction space and show that a loss function in the discretized space approximates our original loss well, owing to its Lipschitz continuity. We then show how the fair regression problem in this discretized space can be turned into a constrained classification problem, which we solve via reduction.

4.1. Discretization

We discretize both arguments of the loss function ℓ. Let N denote the size of the discretization grid for the second argument, let $\alpha = 1/N$ denote its granularity, and let $\mathcal{Z} = \{j\alpha : j = 1, \ldots, N\}$ denote the grid itself. Let $\tilde{\mathcal{Y}}$ be the $\frac{\alpha}{2}$-cover of $\mathcal{Y}$, i.e., $\tilde{\mathcal{Y}} \subseteq \mathcal{Y}$ such that: (1) for any $y \in \mathcal{Y}$ there exists $\tilde{y} \in \tilde{\mathcal{Y}}$ such that $|y - \tilde{y}| \leq \frac{\alpha}{2}$, and (2) for any $\tilde{y}, \tilde{y}' \in \tilde{\mathcal{Y}}$, we have $|\tilde{y} - \tilde{y}'| > \frac{\alpha}{2}$. Proceeding left-to-right within $\mathcal{Y}$, it is always possible to construct $\tilde{\mathcal{Y}}$ such that $|\tilde{\mathcal{Y}}| \leq 2N$. We define the discretized loss as a piece-wise constant approximation of ℓ:

$$\ell_\alpha(y, u) \coloneqq \ell\left(\underline{y}, \lfloor u \rfloor_\alpha + \tfrac{\alpha}{2}\right) \quad (5)$$

where $\underline{y}$ is the smallest $\tilde{y} \in \tilde{\mathcal{Y}}$ such that $|y - \tilde{y}| \leq \frac{\alpha}{2}$, and $\lfloor u \rfloor_\alpha$ rounds down u to the nearest integer multiple of α. We use the convention $\ell(y, u) = \ell(y, 1)$ for $u \geq 1$. Owing to the Lipschitz continuity of ℓ, it follows that

$$\big|\ell(y, u) - \ell_\alpha(y, u)\big| \leq \alpha. \quad (6)$$

Thus, for suitably small α, or equivalently large N, ℓ_α provides a close approximation to the original loss function.

Let $\text{loss}_\alpha(f) \coloneqq \mathbb{E}[\ell_\alpha(Y, f(X))]$ denote the expected discretized loss. When optimizing this loss, it suffices to consider rounded-down variants of predictors. Specifically, for $f \in \mathcal{F}$, let $\underline{f}(x) = \lfloor f(x) \rfloor_\alpha$ denote its rounded-down version. Then, by the definition of ℓ_α, $\text{loss}_\alpha(f) = \text{loss}_\alpha(\underline{f})$. The advantage of rounded-down predictors is that to guarantee that they satisfy SP, it suffices to consider the fairness constraints $\gamma_{a,z}(\underline{f}) \le \varepsilon_a$ across z taken from the discretization grid $\mathcal{Z}$. This is because for any $z \in [0,1]$,

$$\begin{aligned}\gamma_{a,z}(\underline{f}) &= \mathbb{P}[\underline{f}(X) \ge z \mid A = a] - \mathbb{P}[\underline{f}(X) \ge z] \\ &= \mathbb{P}[\underline{f}(X) \ge \bar{z} \mid A = a] - \mathbb{P}[\underline{f}(X) \ge \bar{z}], \quad (7)\end{aligned}$$

where $\bar{z} = \lceil z \rceil_\alpha$ is the value of z rounded up to the nearest integer multiple of α. This allows us to replace the uncountable set of constraints indexed by $z \in [0,1]$ with the finite set indexed by $z \in \mathcal{Z}$. Thus, denoting $\underline{\mathcal{F}} = \{\underline{f} : f \in \mathcal{F}\}$, we have argued that the solution of (3), can be approximated by

$$\min_{Q \in \Delta(\underline{\mathcal{F}})} \text{loss}_\alpha(Q) \text{ s.t. } \left|\gamma_{a,z}(Q)\right| \le \varepsilon_a \ \ \forall a \in \mathcal{A}, z \in \mathcal{Z}. \quad (8)$$

Theorem 1. *Let $Q^\star$ be any feasible point of* (3) *and $Q^\star$ be the solution of* (8)*. Then* $\text{loss}(Q^\star) \le \text{loss}(Q^\star) + \alpha$ *and* $|\gamma_{a,z}(Q^\star)| \le \varepsilon_a$ *for all $a \in \mathcal{A}$, $z \in [0,1]$.*

4.2. Reduction to Constrained Classification

We next show that (8) can be rewritten as a constrained classification problem for the family of classifiers $\mathcal{H} = \{h_f : f \in \mathcal{F}\}$ defined in Section 3.

To turn regression loss ℓ into a cost-sensitive loss, we introduce the function

$$c(y,z) \coloneqq N\Big(\ell\left(y, z + \tfrac{\alpha}{2}\right) - \ell\left(y, z - \tfrac{\alpha}{2}\right)\Big), \quad (9)$$

which takes values in $[-1,1]$, because ℓ is 1-Lipschitz and $\alpha = 1/N$. We also extend the $\gamma_{a,z}$ notation to h_f:

$$\gamma_{a,z}(h_f) = \mathbb{E}[h_f(X,z) | A = a] - \mathbb{E}[h_f(X,z)].$$

Now given a distribution D over (X, A, Y), we define a distribution D' over (X', A, C) that additionally samples $Z \in \mathcal{Z}$ uniformly at random and sets $X' = (X, Z)$ and $C = c(\underline{Y}, Z)$. Defining $\text{cost}(h_f) \coloneqq \mathbb{E}_{D'}[C h_f(X')]$, we have the following useful lemma.

Lemma 1. *Given any distribution D over (X, A, Y) and any $f \in \mathcal{F}$, the cost and constraints satisfy* $\text{cost}(h_f) = \text{loss}_\alpha(\underline{f}) + c_0$*, where c_0 is independent of f, and $\gamma_{a,z}(h_f) = \gamma_{a,z}(\underline{f})$ for all $a \in \mathcal{A}$, $z \in \mathcal{Z}$.*

By linearity of expectation, the lemma implies analogous equalities also for distributions over f. Thus, in problem (8), we can replace the optimization over $Q \in \Delta(\underline{\mathcal{F}})$ with $Q \in \Delta(\mathcal{H})$. Notice that while we started from discretized regressors in problem (8), Lemma 1 allows us to work with the full classifier family $\{h_f : f \in \mathcal{F}\}$, which is important as we typically only have computational oracles for non-discretized classes $\mathcal{H}$ and $\mathcal{F}$. We next show how to solve an empirical version of this classification problem.

4.3. Algorithm and Generalization Bounds

Let $\widehat{\mathbb{E}}$ denote the empirical distribution over the data (X_i, A_i, Y_i) and let $\mathbb{E}_Z$ denote a uniform distribution among the values in $\mathcal{Z}$. Then define the empirical versions of the cost and constraints:

$$\begin{aligned}\widehat{\text{cost}}(h_f) &= \widehat{\mathbb{E}}\Big[\mathbb{E}_Z\big[c(\underline{Y}, Z) h_f(X, Z)\big]\Big] \quad (10)\\ \widehat{\gamma}_{a,z}(h_f) &= \widehat{\mathbb{E}}\big[h_f(X,z) \bigm| A = a\big] - \widehat{\mathbb{E}}\big[h_f(X,z)\big].\end{aligned}$$

We are interested in the following empirical optimization problem, which is, according to Lemma 1, an empirical approximation of the original problem (8):

$$\min_{Q \in \Delta(\mathcal{H})} \widehat{\text{cost}}(Q) \text{ s.t. } \left|\widehat{\gamma}_{a,z}(Q)\right| \le \widehat{\varepsilon}_a \ \ \forall a \in \mathcal{A}, z \in \mathcal{Z}. \quad (11)$$

The slacks $\widehat{\varepsilon}_a$ are slightly larger than ε_a to compensate for finite-sample errors in measuring constraint violations (more on that below). This problem is a special case of that studied by Agarwal et al. (2018) with a key difference. Since the distribution of Z is known, we can take expectation according to Z rather than a sample, which leads to substantially better estimates of constraint violations. Thus, our objective uses a product of an empirical distribution over (X, A, Y) with the uniform distribution over Z rather than an i.i.d. sample as assumed by Agarwal et al.. However, their algorithm and generalization bounds still apply (as we show in our proofs).

The algorithm begins by forming the Lagrangian with the primal variable $Q \in \Delta(\mathcal{H})$ and the dual variable $\boldsymbol{\lambda}$ with components $\lambda^+_{a,z}, \lambda^-_{a,z} \in \mathbb{R}_+$, corresponding to the constraints $\widehat{\gamma}_{a,z}(Q) \le \widehat{\varepsilon}_a$ and $\widehat{\gamma}_{a,z}(Q) \ge -\widehat{\varepsilon}_a$:

$$\begin{aligned}L(Q, \boldsymbol{\lambda}) = \widehat{\text{cost}}(Q) + \sum_{a,z} \Big[&\lambda^+_{a,z}\big(\widehat{\gamma}_{a,z}(Q) - \widehat{\varepsilon}_a\big) \\ &+ \lambda^-_{a,z}\big(-\widehat{\gamma}_{a,z}(Q) - \widehat{\varepsilon}_a\big)\Big].\end{aligned}$$

It solves the saddle-point problem $\min_Q \max_{\boldsymbol{\lambda}} L(Q, \boldsymbol{\lambda})$ over $Q \in \Delta(\mathcal{H})$ and $\boldsymbol{\lambda} \ge \mathbf{0}$, $\|\boldsymbol{\lambda}\|_1 \le B$, by treating it as a two-player zero-sum game (see Algorithm 1 for details).

We bound the suboptimality and fairness of the returned solution largely following their analysis. Let $R_n(\mathcal{H})$ denote the Rademacher complexity of $\mathcal{H}$ (see Eq. 17 in Appendix C). To state the bounds, recall an assumption from their paper on the setting of the empirical slacks $\widehat{\varepsilon}_a$:

Assumption 1. *There exist $C, C' > 0$ and $\beta \le 1/2$ such that $R_n(\mathcal{H}) \le C n^{-\beta}$ and $\widehat{\varepsilon}_a = \varepsilon_a + C' n_a^{-\beta}$, where n_a is the number of samples with $A = a$.*

Under this assumption, we obtain the following guarantees.[4]

Theorem 2. *Let Assumption 1 hold for $C' \ge 2C + 2 + \sqrt{2\ln(4|\mathcal{A}|N/\delta)}$, where $\delta > 0$. Let $Q^\star$ be any feasible distribution for the fair regression problem* (3)*. Then Algorithm 1 with $\nu \propto n^{-\beta}$, $B \propto n^\beta$, and $N \propto n^\beta$ terminates in $O\left(n^{4\beta} \ln(n^\beta |\mathcal{A}|)\right)$ iterations and returns $\widehat{Q}$, which, when*

[4] The notation $\widetilde{O}(\cdot)$ suppresses polynomial dependence on $\ln n$, $\ln|\mathcal{A}|$, and $\ln(1/\delta)$.

Algorithm 1 Fair regression with statistical parity

Input: training examples $\{(X_i, Y_i, A_i)\}_{i=1}^n$, slacks $\widehat{\varepsilon}_a \in [0,1]$, bound B, threshold ν

Define best-response functions:
$\textsc{Best}_h(\boldsymbol{\lambda}) \coloneqq \arg\min_{h_f \in \mathcal{H}} L(h_f, \boldsymbol{\lambda})$
$\textsc{Best}_{\boldsymbol{\lambda}}(Q) \coloneqq \arg\max_{\boldsymbol{\lambda} \geq 0,\, \|\boldsymbol{\lambda}\|_1 \leq B} L(Q, \boldsymbol{\lambda})$

Set learning rate $\eta = \nu/(8B)$
Set $\boldsymbol{\theta}_1^+ = \boldsymbol{\theta}_1^- = \mathbf{0} \in \mathbb{R}^{|\mathcal{A}|N}$
for $t = 1, 2, \ldots$ **do**
  // Compute $\boldsymbol{\lambda}_t$ from $\boldsymbol{\theta}_t$ and find the best response h_t
  for all a, z and $\sigma \in \{+,-\}$ **do**
    $\lambda_{t,a,z}^{\sigma} \leftarrow B \frac{\exp\{\theta_{t,a,z}^{\sigma}\}}{1+\sum_{a,z}\left[\exp\{\theta_{t,a,z}^{+}\}+\exp\{\theta_{t,a,z}^{-}\}\right]}$
  end for
  $h_t \leftarrow \textsc{Best}_h(\boldsymbol{\lambda}_t)$
  // Calculate the current approximate saddle point
  $\widehat{Q}_t \leftarrow \frac{1}{t}\sum_{t'=1}^{t} h_{t'}, \quad \widehat{\boldsymbol{\lambda}}_t \leftarrow \frac{1}{t}\sum_{t'=1}^{t} \boldsymbol{\lambda}_{t'}$
  // Check the suboptimality of $(\widehat{Q}_t, \widehat{\boldsymbol{\lambda}}_t)$
  $\overline{\nu} \leftarrow L\big(\widehat{Q}_t, \textsc{Best}_{\boldsymbol{\lambda}}(\widehat{Q}_t)\big) \;-\; L(\widehat{Q}_t, \widehat{\boldsymbol{\lambda}}_t)$
  $\underline{\nu} \leftarrow L(\widehat{Q}_t, \widehat{\boldsymbol{\lambda}}_t) \;-\; L\big(\textsc{Best}_h(\widehat{\boldsymbol{\lambda}}_t), \widehat{\boldsymbol{\lambda}}_t\big)$
  if $\max\{\overline{\nu}, \underline{\nu}\} \leq \nu$ **then return** $\widehat{Q}_t$
  // Apply the exponentiated-gradient update
  Set $\boldsymbol{\theta}_{t+1}^{\sigma} = \boldsymbol{\theta}_t^{\sigma} + \sigma\eta\widehat{\boldsymbol{\gamma}}(h_t) - \eta\widehat{\boldsymbol{\varepsilon}}$, for $\sigma \in \{+,-\}$
end for

viewed as a distribution over $\underline{\mathcal{F}}$, satisfies with probability at least 1-δ,

$$\mathrm{loss}(\widehat{Q}) \leq \mathrm{loss}(Q^\star) + \widetilde{O}(n^{-\beta})$$
$$\left|\gamma_{a,z}(\widehat{Q})\right| \leq \varepsilon_a + \widetilde{O}(n_a^{-\beta}) \quad \textit{for all } a \in \mathcal{A},\, z \in [0,1].$$

Note that the bounds grow with the Rademacher complexity of $\mathcal{H}$, rather than the complexity of the regressor class $\mathcal{F}$. Since $\left|\mathbb{E}_Z[h_f(X,Z)] - f(X)\right| \leq \alpha$, it can be shown that $R_n(\mathcal{H}) \geq R_n(\mathcal{F}) - \alpha$, meaning that the classifiers induce a more complex class. The bound on $\mathrm{loss}(\widehat{Q})$ in Theorem 2 can be stated in terms of the tighter $R_n(\mathcal{F})$, but the constraints still deviate by $R_n(\mathcal{H})$, which we believe is unavoidable. However, if $\mathcal{F}$ has a bounded pseudo-dimension, which always equals the VC dimension of $\mathcal{H}$ (Anthony & Bartlett, 2009), then the pseudo-dimension can be used to bound $R_n(\mathcal{H})$ (see Theorem 6 of Bartlett & Mendelson, 2002).

4.4. Efficient Implementation of Algorithm 1

It is not too difficult to show that each iteration of Algorithm 1 can be implemented in time $O(n \log n + |\mathcal{A}|N)$ plus the complexity of two calls to $\textsc{Best}_h$, on which we focus here, while deferring the remaining details to Appendix F.

Reduction to cost-sensitive classification. We first show how to implement $\textsc{Best}_h$ using the CS oracle. Letting $\lambda_{a,j} = \lambda_{a,j}^+ - \lambda_{a,j}^-$, and dropping terms independent of h, the minimization over h only needs to be over $\widehat{\mathrm{cost}}(h) + \sum_{a,z} \lambda_{a,z}\widehat{\gamma}_{a,z}(h)$. The first term is already defined as CS error with respect to $\widehat{\mathbb{E}}$ and $\mathbb{E}_Z$ (see Eq. 10). Let $p_a \coloneqq \widehat{\mathbb{P}}[A = a]$. Then we show in Appendix F that the minimization over $L(h, \boldsymbol{\lambda})$ is equivalent to minimizing

$$\widehat{\mathrm{cost}}(h) + \sum_{a,z} \lambda_{a,z}\widehat{\gamma}_{a,z}(h) \tag{12}$$
$$= \widehat{\mathbb{E}}\left[\mathbb{E}_Z\left[\underbrace{\left(c(\underline{Y}, Z) + \frac{N\lambda_{A,Z}}{p_A} - \sum_a N\lambda_{a,Z}\right)}_{c_{\boldsymbol{\lambda}}(\underline{Y}, A, Z)} h(X, Z)\right]\right].$$

This corresponds to a CS problem with nN instances $\{(X'_{i,z}, C_{i,z})\}_{i \leq n,\, z \in \mathcal{Z}}$ defined as

$$X'_{i,z} = (X_i, z), \quad C_{i,z} = c_{\boldsymbol{\lambda}}(\underline{Y}_i, A_i, z).$$

The sum $\sum_a N\lambda_{a,Z}$ in the definition of $c_{\boldsymbol{\lambda}}$ can be precalculated once for each value of Z in the overall time $O(N|\mathcal{A}|)$. After that the construction of the dataset takes time $O(nN)$.

Based on Assumption 1 and Theorem 2, we expect $N \propto n^\beta$, so this reduction to cost-sensitive classification takes time $\Omega(n^{1+\beta})$ and creates a dataset of size $\Omega(n^{1+\beta})$. This is substantially larger than the original problem of size n, given the typical value $\beta \approx 1/2$. We next describe two alternatives that run faster and only create datasets of size n.

Reduction to least-squares regression. The main overhead in the CS reduction above comes from the summation over $z \in \mathcal{Z}$, implicit in the expectation over Z in Eq. (12). In order to eliminate this overhead, suppose we have access to a function $g_{\boldsymbol{\lambda}}$ such that

$$g_{\boldsymbol{\lambda}}(\tilde{Y}, A, f(X)) = \frac{1}{N}\sum_{z \in \mathcal{Z}} c_{\boldsymbol{\lambda}}(\tilde{Y}, A, z) h_f(X, z)$$

for any $\tilde{Y} \in \tilde{\mathcal{Y}}$, $A \in \mathcal{A}$, $X \in \mathcal{X}$ and $f \in \mathcal{F}$. In Appendix F.2, we show how to precalculate $g_{\boldsymbol{\lambda}}(\cdot,\cdot,\cdot)$ in time $O(|\mathcal{A}||\tilde{\mathcal{Y}}|N)$. Then minimizing Eq. (12) over h is equivalent to finding $\min_{f \in \mathcal{F}} \sum_{i=1}^n g_{\boldsymbol{\lambda}}(\underline{Y}_i, A_i, f(X_i))$. We heuristically solve this problem by calling a least-squares oracle on a dataset of size n. To this end, we replace $g_{\boldsymbol{\lambda}}$ by the square loss with respect to specific targets U_i (different from Y_i). As targets, we choose the minima of $g_{\boldsymbol{\lambda}}$ for each fixed $\underline{Y}_i$ and A_i, that is, $U_i \in \arg\min_{u \in [0,1]} g_{\boldsymbol{\lambda}}(\underline{Y}_i, A_i, u)$. We seek to solve

$$\min_{f \in \mathcal{F}} \sum_{i=1}^n (U_i - f(X_i))^2.$$

To obtain the values U_i we first calculate $c_{\boldsymbol{\lambda}}(\tilde{y}, a, z)$ across all $\tilde{y}$, a, and z in the overall time $O(|\mathcal{A}||\tilde{\mathcal{Y}}|N)$. Then, using the definition of $g_{\boldsymbol{\lambda}}$, the minimizer of $g_{\boldsymbol{\lambda}}(\tilde{y}, a, u)$ over u can be found in time $O(N)$ for each pair of $\tilde{y}$ and a, so all the minimizers can be precalculated in time $O(|\mathcal{A}||\tilde{\mathcal{Y}}|N)$. Thus, preparing the dataset for the least-squares reduction takes time $O(|\mathcal{A}||\tilde{\mathcal{Y}}|N)$ and the resulting regression dataset is of size n. Since $|\tilde{\mathcal{Y}}| = O(N)$, and $N \propto n^\beta$. The running time of the reduction is $O(n \log n + |\mathcal{A}|n^{2\beta})$, which is

substantially faster than $\Omega(n^{1+\beta})$ for typical $\beta \approx 1/2$.

Reduction to risk minimization under ℓ. A similar heuristic as in the least-squares reduction can be used to reduce $\textsc{Best}_h$ to risk minimization under any loss $\ell(y, u)$ that is convex in u. The complexity of this reduction is identical to that of the least-squares reduction, but the resulting risk minimization problem might be better aligned with the original problem, yielding a potentially superior oracle as we will see in the experiments. (See Appendix F.3 for details.)

5. Fair Regression with Bounded Group Loss

We now turn attention to our second notion of fairness. We show how to reduce fair regression with bounded group loss to loss minimization under ℓ without the fairness constraints.

The approach still follows the scheme of Agarwal et al. (2018), but thanks to the matched loss function between the objective and the constraints, fair regression can be reduced directly to regression, without the need for discretization. We first replace the problem (4) by its empirical version

$$\min_{Q \in \Delta(\mathcal{F})} \widehat{\text{loss}}(Q) \;\text{ s.t. }\; \widehat{\gamma}_a^{\text{BGL}}(Q) \le \widehat{\zeta}_a \;\; \forall a \in \mathcal{A}. \tag{13}$$

We then form the Lagrangian with the primal variable Q and the dual variable $\boldsymbol{\lambda}$ with components $\lambda_a \in \mathbb{R}_+$ corresponding to the constraints $\widehat{\gamma}_a^{\text{BGL}}(Q) \le \widehat{\zeta}_a$:

$$L^{\text{BGL}}(Q, \boldsymbol{\lambda}) = \widehat{\text{loss}}(Q) + \sum_a \lambda_a \Big(\widehat{\gamma}_a^{\text{BGL}}(Q) - \widehat{\zeta}_a \Big).$$

We give a detailed pseudocode for our approach in Algorithm 2 in Appendix D, and describe the main differences from Algorithm 1 here. As before, the algorithm alternates between exponentiated gradient updates on $\boldsymbol{\lambda}$ and best responses for Q to compute an approximate saddle point:

$$\min_{Q \in \Delta} \max_{\boldsymbol{\lambda} \ge \mathbf{0},\, \|\boldsymbol{\lambda}\|_1 \le B} L^{\text{BGL}}(Q, \boldsymbol{\lambda}). \tag{14}$$

The saddle point always exists. However, unlike the fair regression problem under SP, the fair regression problem under BGL, i.e., Eq. (13), might be infeasible. Therefore, Algorithm 2 explicitly checks whether the distribution $\widehat{Q}$ that it finds satisfies constraints of Eq. (13).

The other main difference between Algorithms 1 and 2 is in the computation of the best response f to any given $\boldsymbol{\lambda}$, which requires solving the problem

$$\min_{f \in \mathcal{F}} \Big[\widehat{\text{loss}}(f) + \sum_a \lambda_a \widehat{\gamma}_a^{\text{BGL}}(f) \Big].$$

Denoting by n_a the number of samples with $A_i = a$, this minimization can be written as

$$\min_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)) + \sum_a \frac{\lambda_a}{n_a} \sum_{i:\, A_i = a} \ell(Y_i, f(X_i)) \right],$$

which can be solved using one call to the weighted risk-minimization oracle.

We finish this section with the optimality and fairness guarantees for $\widehat{Q}$ returned by Algorithm 2. We assume that $\widehat{\zeta}_a$ are set according to the Rademacher complexity of $\mathcal{F}$:

Assumption 2. *There exist $C, C' > 0$ and $\omega \le 1/2$ such that $R_n(\mathcal{F}) \le C n^{-\omega}$ and $\widehat{\zeta}_a = \zeta_a + C' n_a^{-\omega}$, where n_a is the number of samples where $A = a$.*

Theorem 3. *Let Assumption 2 hold for $C' \ge 4C + 2 + \sqrt{2\ln(4|\mathcal{A}|/\delta)}$, where $\delta > 0$. Then Algorithm 2 with $\nu \propto n^{-\omega}$ and $B \propto n^{\omega}$ terminates in $O(n^{4\omega} \ln |\mathcal{A}|)$ iterations and returns $\widehat{Q}$ such that, with probability at least 1-δ, one of the following holds:*

1. *$\widehat{Q} \ne$ null and, for any $Q^\star$ feasible in problem* (4)*,*
$$\text{loss}(\widehat{Q}) \le \text{loss}(Q^\star) + \widetilde{O}(n^{-\omega})$$
$$\gamma_a^{\text{BGL}}(\widehat{Q}) \le \zeta_a + \widetilde{O}(n_a^{-\omega}) \quad \textit{for all } a \in \mathcal{A}.$$
2. *$\widehat{Q} =$ null and problem* (4) *is infeasible.*

6. Experiments

We evaluate our method on the tasks of least-squares regression and logistic regression under statistical parity. We use the following three datasets:

Adult: The adult income dataset (Lichman, 2013) has 48,842 examples. The task is to predict the probability that an individual makes more than \$50k per year via logistic loss minimization, with gender as the protected attribute.

Law school: Law School Admissions Council's National Longitudinal Bar Passage Study (Wightman, 1998) has 20,649 examples. The task is to predict a student's GPA (normalized to $[0, 1]$) via square loss minimization, with race as the protected attribute (white versus non-white).

Communities & crime: The dataset contains socio-economic, law enforcement, and crime data about communities in the US (Redmond & Baveja, 2002) with 1,994 examples. The task is to predict the number of violent crimes per 100,000 population (normalized to $[0, 1]$) via square loss minimization, with race as the protected attribute (whether the majority population of the community is white).

For the two larger datasets (*adult* and *law school*), we also created smaller (subsampled) versions by picking random 2,000 points. Thus we ended up with a total of five datasets, and split each into 50% for training and 50% for testing.

We ran Algorithm 1 on each training set over a range of constraint slack values $\hat{\varepsilon}$, with a fixed discretization grid of size 40: $\mathcal{Z} = \{1/40, 2/40, \ldots, 1\}$. Among the solutions for different $\hat{\varepsilon}$, we selected the ones on the Pareto front based on their training losses and SP disparity $\max_{a,z}\{\hat{\gamma}_{a,z}\}$. We then evaluated the selected predictors on the test set, and show the resulting Pareto front in Figure 1.

We ran our algorithm with the three types of reductions

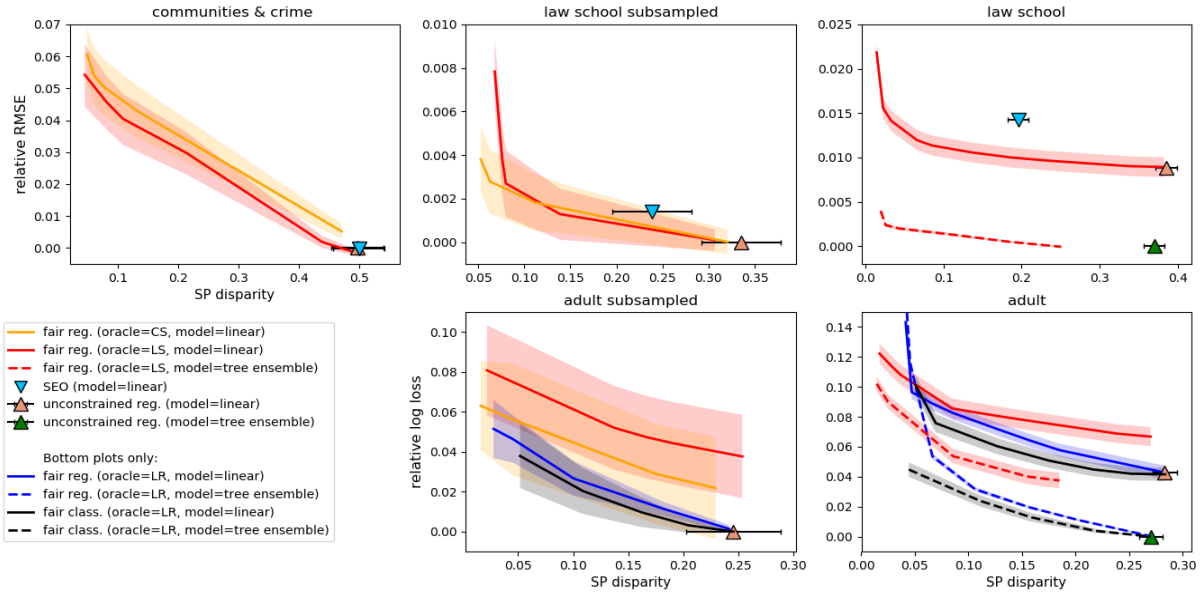

Figure 1. Relative test loss versus the worst constraint violation with respect to SP. Relative losses are computed by subtracting the smallest baseline loss from the actual loss. For our algorithm and fair classification we plot the convex envelope of the predictors obtained on training data at various accuracy–fairness tradeoffs. We show 95% confidence bands for the relative loss of our method and fair classification, and also show 95% confidence intervals for constraint violation (the same for all methods). Our method dominates or matches the baselines up to statistical uncertainty on all datasets except *adult*, where fair classification is slightly better.

from Section 4.4: reductions to cost-sensitive (CS) oracles, least-squares (LS) oracles, and logistic-loss minimization (LR) oracles. Our CS oracle sought the linear model minimizing weighted hinge-loss (as a surrogate for weighted classification error). Because of unfavorable scaling of the cost-sensitive problem sizes (see Section 4.4), we only ran the CS oracle on the three small datasets. We considered two variants of LS and LR oracles: linear learners from scikit-learn (Pedregosa et al., 2011), and tree ensembles from XGBoost (Chen & Guestrin, 2016). Tree ensembles heavily overfitted smaller datasets, so we only show their performance on two larger datasets. We only used LR oracles when the target loss was logistic, whereas we used LS oracles across all datasets.

In addition to our algorithm, we also evaluated regression without any fairness constraints, and two baselines from the fair classification and fair regression literature.

On the three datasets where the task was least-squares regression, we evaluated the *full substantive equality of opportunity* (SEO) estimate of Johnson et al. (2016). It can be obtained in a closed form by solving for the linear model that minimizes least-squares error while having zero correlation with the protected attribute. In contrast, our method seeks to limit not just correlation, but statistical dependence.

On the two datasets where the task was logistic regression, we ran the *fair classification* (FC) reduction of Agarwal et al. (2018) with the same LR oracles as in our algorithm. For this choice of oracles, the classifiers returned by FC are implemented by logistic models and return real-valued scores, which we evaluated. We ran FC across a range of trade-offs between classification accuracy and statistical parity (in the classification sense) and show the resulting Pareto front. Note that FC only enforces statistical parity (SP) when the scores are thresholded at zero, whereas our method enforces SP across all thresholds.

In Figure 1, we see that all of our reductions are able to significantly reduce disparity, without strongly impacting the overall loss. On *communities & crime*, there is a more substantial accuracy–fairness tradeoff, which can be used as a starting point to diagnose the data quality for the two racial subgroups. Our methods dominate SEO in least-squares tasks, but are slightly worse than FC in logistic regression. The difference is statistically significant only on *adult*, where it points to the limitations of our LS and LR reduction heuristics. However, for the most part, LR and LS reductions achieve tradeoffs on par with the CS reduction, and are substantially faster to run (see Appendix G). The results on *adult* and *adult subsampled* suggest that reducing to a matching loss is preferable over reducing to another loss.

In summary, we have shown that our scheme efficiently handles a range of losses and regressor classes and, where possible, diminishes disparity while maintaining the overall accuracy. The emergence of FC as a strong baseline for logistic regression suggests that our regression-based reduction heuristics can be further improved, which we leave open for future research.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *ICML*, 2018.

Alabi, D., Immorlica, N., and Kalai, A. Unleashing linear optimizers for group-fair learning and optimization. In *Conference on Learning Theory*, 2018.

Aliprantis, C. and Border, K. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer Berlin Heidelberg, 2006.

Anthony, M. and Bartlett, P. L. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

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

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in criminal justice risk assessments: The state of the art. arXiv:1703.09207, 2017.

Boucheron, S., Bousquet, O., and Lugosi, G. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pp. 77–91, 2018.

Calders, T., Karim, A., Kamiran, F., Ali, W., and Zhang, X. Controlling attribute effect in linear regression. In *2013 IEEE 13th International Conference on Data Mining*, pp. 71–80, 2013.

Chen, I., Johansson, F. D., and Sontag, D. Why is my classifier discriminatory? In *NeurIPS*, 2018.

Chen, T. and Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, 2016.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, Special Issue on Social and Technical Trade-Offs, 2017.

Corbett-Davies, S. and Goel, S. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv, 2018.

Dieterich, W., Mendoza, C., and Brennan, T. Compas risk scales: Demonstrating accuracy equity and predictive parity. 2016.

Fukuchi, K., Sakuma, J., and Kamishima, T. Prediction with model-based neutrality. In *Machine Learning and Knowledge Discovery in Databases—European Conference, ECML PKDD 2013*, pp. 499–514, 2013.

Gurobi Optimization, L. Gurobi optimizer reference manual, 2018. URL `http://www.gurobi.com`.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Neural Information Processing Systems (NIPS)*, 2016.

Holzer, H. J. and Neumark, D. Affirmative action: What do we know? *Journal of Policy Analysis and Management*, 25(2):463–490, 2006.

Johnson, K. D., Foster, D. P., and Stine, R. A. Impartial predictive modeling: Ensuring fairness in arbitrary models. arXiv:1608.00528, 2016.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases—European Conference, ECML PKDD 2012*, pp. 35–50, 2012.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, 2017.

Komiyama, J., Takeda, A., Honda, J., and Shimao, H. Nonconvex optimization for regression with fairness constraints. In *International Conference on Machine Learning*, pp. 2742–2751, 2018.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.

Lehmann, E. L. and Romano, J. P. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

Lichman, M. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

Lowenkamp, C., L Johnson, J., Holsinger, A., Vanbenschoten, S., and R Robinson, C. The federal post conviction risk assessment (pcra): A construction and validation study. *Psychological services*, 10, 2012.

Northpointe. Compas risk & need assessment system: Selected questions posed by inquiring agencies. 2010.

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

Pérez-Suay, A., Laparra, V., Mateo-García, G., Muñoz-Marí, J., Gómez-Chova, L., and Camps-Valls, G. Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 339–355. Springer, 2017.

Podesta, J., Pritzker, P., Moniz, E. J., Holdren, J., and Zients, J. Big data: Seizing opportunities and preserving values. 2014.

Redmond, M. and Baveja, A. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research 14*, 2002.

Selbst, A. D., Friedler, S., Venkatasubramanian, S., Vertesi, J., et al. Fairness and abstraction in sociotechnical systems. In *ACM Conference on Fairness, Accountability, and Transparency (FAT*)*, 2018.

US Federal Reserve. Report to the congress on credit scoring and its effects on the availability and affordability of credit. 2007.

Veale, M., Van Kleek, M., and Binns, R. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

Waters, A. and Miikkulainen, R. Grade: Machine learning support for graduate admissions. *AI Magazine*, 35(1):64, 2014.

Wightman, L. LSAC National Longitudinal Bar Passage Study, 1998.