A Possibility in Algorithmic Fairness: Calibrated Scores for Fair Classifications

Claire Lazar Reich
MIT Statistics & Economics
clazar@mit.edu

Suhas Vijaykumar
MIT Statistics & Economics
suhasv@mit.edu

Abstract

Calibration and equal error rates are fundamental criteria of algorithmic fairness that have been shown to conflict with one another. This paper proves that they can be satisfied simultaneously in settings where decision-makers use risk scores to assign binary treatments. In particular, we derive necessary and sufficient conditions for the existence of calibrated scores that yield classifications achieving equal error rates. We then present an algorithm that searches for the most informative score subject to both calibration and minimal error rate disparity. Applied to a real criminal justice risk assessment, we show that our method can eliminate error disparities while maintaining calibration. In a separate application to credit lending, the procedure provides a solution that is more fair and profitable than a common alternative that omits sensitive features.

1 Introduction

Today's algorithms reach deep into decisions that guide our lives, from loan approvals to medical treatments to foster care placements. While they hold the promise of driving social advancements, making them fair has proven to be challenging both in practice and in theory. Recent work has shown that even when a dataset is free from bias, an algorithm trained on it will face significant fairness tradeoffs as long as groups represented in the data have different average outcomes [14, 5, 4, 8, 13]. These fairness impossibility results have underscored the need to target certain criteria of fairness at the expense of others.

In contrast, this paper presents a natural setting in which it is possible to reconcile two important but conflicting notions of fairness: calibration and equal error rates. In influential recent work, these two criteria were proven to be mutually incompatible when both are applied to a *risk score* [14, 22] and when both are applied to a *classifier* [5], suggesting that they may be incompatible altogether [1, 7].

We relax the mathematical tension between these two fairness criteria by enforcing *calibration on the score* and *equal error rates on the corresponding classifier*. In particular, we show that it is possible to design calibrated scores that yield equal error rate classifications at group-blind cutoffs, and we provide a method to do so in practice. Modern risk assessments that output scores and classification recommendations can use the method to satisfy both fairness criteria.

In addition, our framework and method can be applied toward the goal of achieving equal error rates in human decisions. We consider settings in which a risk score is provided to an accuracy-maximizing agent, such as a lender, who then uses it to assign binary treatments, such as loan approvals and denials. We show how to deliver calibrated scores that lead the agent to make the most accurate classifications that satisfy equal error rates. As a consequence, a creditworthy applicant's probability of being granted a loan will not depend on their group affiliation [11].

A key theme in our analysis is that data richness complements our fairness criteria. On the theoretical side, we illustrate that the set of attainable error rates grows with the informativeness of the data. In an empirical application to credit lending, we compare our method to the common practice of omitting sensitive features from training data and show that it simultaneously achieves higher accuracy and lower error rate disparity. These results on calibration and equal error rates contribute to growing evidence that omitting sensitive features can be misguided [9, 10, 6, 16, 13].

Our results proceed as follows. In Section 2, we prove that it is possible to construct calibrated scores that lead to equal error rate classifications and we precisely characterize when such scores exist. In Section 3, we propose an algorithm that produces the most accurate possible score satisfying the fairness criteria and minimizing the decision-maker's errors. Finally, in Section 4, we apply our proposed method to two empirical settings. We first assess its performance in helping a lender screen loan applicants. We then apply the method to the COMPAS criminal risk assessment tool, where we show that our procedure can eliminate error rate imbalances in risk classifications while preserving calibration of scores.

2 Theoretical Results

2.1 Formal Setting

Let us consider a triple (Y, X, A) on a common probability space $\mathbb{P}$, where $Y \in \{0, 1\}$ is an outcome variable, $X \in \mathbb{R}^d$ is a vector of features, and $A \in \{H, L\}$ is a protected attribute differentiating two groups with unequal base rates of the outcome,

$$\mathbb{E}[Y|A = L] < \mathbb{E}[Y|A = H]. \tag{1}$$

Our goal is to estimate a score function $\hat{p} \equiv \hat{p}(X, A) \in [0, 1]$ that predicts Y with maximum accuracy subject to the fairness constraints of calibration and equal error rates. Specifically, we hand $\hat{p}$ to a decision-maker tasked with selecting classifications $\hat{y} \in \{0, 1\}$ that minimize the loss function

$$\ell(\hat{y}, y) = \begin{cases} 0 & y = \hat{y} \\ 1 & y > \hat{y} \\ k & y < \hat{y}, \end{cases} \tag{2}$$

where $k > 0$ is the relative cost of false positive classifications. Note that any loss function that is minimized when $y = \hat{y}$ is equivalent to ℓ after an affine transformation.

Let us suppose the decision-maker can observe group affiliation A in addition to $\hat{p}$. To ensure that classifications are based only on $\hat{p}$ and not on A, we constrain $\hat{p}$ to satisfy *calibration within groups*,

$$\mathbb{E}[Y|A, \hat{p}] = \mathbb{E}[Y|\hat{p}] = \hat{p}. \tag{3}$$

If (3) holds, the decision-maker's expected loss given $\hat{p}$ and A becomes

$$\mathbb{E}[\ell(Y, \hat{y})|\hat{p}, A] = \hat{p}(1 - \hat{y}) + k(1 - \hat{p})\hat{y}. \tag{4}$$

This expected loss is minimized with a cutoff decision rule that is independent of group affiliation A,

$$\hat{y} = \mathbb{1}\{\hat{p} \geq \bar{p}\}, \tag{5}$$

where the cutoff $\bar{p} = {}^{k}/_{(k+1)}$ is fixed by the decision-maker. Our second fairness condition constrains $\hat{y}$ to satisfy *equal error rates*, ensuring that the classification only depends on the group through the target variable. Following the decision rule (5), we may write this as

$$(\mathbb{1}\{\hat{p} \geq \bar{p}\} \perp\!\!\!\perp A) \mid Y. \tag{6}$$

Our calibration and equal error rate conditions are summarized by (3) and (6), respectively.

2.2 Relation to Impossibility Results

We first introduce a general impossibility result, relate it to previous work, and show where our assumptions diverge to make our proposed criteria satisfiable. The following theorem proves that a *single* algorithmic output Z cannot generally satisfy notions of both calibration and equal error rates.

Theorem 1. Let Y, A, and Z be random variables satisfying the following three conditions.

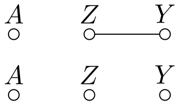
(i) $(Y \perp\!\!\!\perp A) \mid Z$,

(ii) $(Z \perp\!\!\!\perp A) \mid Y$,

(iii) $\mathbb{P}(A = H|Z)$, $\mathbb{P}(Y = 1|A, Z) \in (0, 1)$.

Then A and (Z, Y) must be independent.

Proof. Suppose that (Y, A, Z) satisfy (i) (ii) and (iii). Assumption (iii) implies that the law of (A, Y, Z) is strictly positive. By the Hammersley-Clifford theorem (see e.g. [19]), the conditional independence relations are summarized by a graph on $\{Y, A, Z\}$ where every path from Y to A travels through Z, and every path from A to Z travels through Y. There are only two graphs with this property:



In neither of these graphs does there exist a path from A to (Y, Z), so we conclude that A and (Y, Z) must be independent for (i) (ii) and (iii) to simultaneously hold. □

Note that when A denotes group affiliation and Y denotes outcomes, (i) is a form of calibration and (ii) is a form of the equal error rate condition. Assumption (iii) is a strong form of predictive uncertainty that is generalized in the appendix. Thus the theorem shows that when there is predictive uncertainty and Y depends on A (i.e. when the base rates are unequal), it is impossible for a single Z to satisfy both calibration and equal error rates. For example, letting Z be a classifier recovers the Chouldechova [5] result that (i) equal positive and negative predictive values are unachievable alongside (ii) equal error rates. Meanwhile, letting Z be a risk score shows that (i) calibration is unachievable alongside (ii) a condition that implies balance in the positive and negative class, similar to Kleinberg et al. [14].

Our own setting bypasses the mathematical impossibility described in Theorem 1 by imposing fairness constraints on *two* separate algorithmic outputs rather than one. We require (i) calibration from the scores $\hat{p}$ and (ii) equal error rates from the resulting classifications $\hat{y} = \mathbb{1}\{\hat{p} \geq \bar{p}\}$.

2.3 Necessary and Sufficient Conditions

In this section we characterize exactly when there exists a calibrated $\hat{p}$ that leads to equal error rate classifications $\hat{y}$ at the fixed cutoff $\bar{p}$. Our conditions can be easily checked in a given setting, and they are shown to depend on the informativeness of the features X.

The graphical framework in this section builds on methods developed in Hardt et al. [11]. All the necessary and sufficient conditions will be illustrated in $\mathbb{R}^2$, with true positive rates on the vertical axis and false positive rates on the horizontal. The *feasible region* will be the space in $\mathbb{R}^2$ corresponding to error rates achievable by an equal error rate classifier $\hat{y} = \mathbb{1}\{\hat{p} \geq \bar{p}\}$ where $\hat{p}$ is calibrated.

We first study the entire space of possible equal error rate classifiers, without regard to calibration or the decision-maker's cutoff $\bar{p}$. Then we study the entire space of classifiers that can be based on cutoff rules $\bar{p}$ applied to calibrated scores, without regard to the equal error rate condition. Finally, we assert that the intersection of these two spaces determines fairness feasibility, and we characterize when the intersection is nonempty.

2.3.1 Classifiers Satisfying Equal Error Rates

We wish to identify the entire space of error rates in $\mathbb{R}^2$ achievable by classifiers with equal error rates. Hardt et al. [11] succeeded in doing so, and we review and adapt their results in this subsection. To lay the groundwork for the geometric reasoning to follow, we first denote the group A false positive rate and true positive rate associated with a given classifier $\hat{y}$ as a point in $\mathbb{R}^2$,

$$\alpha(\hat{y}, A) = \Big(\mathbb{P}(\hat{y} = 1|Y = 0, A), \, \mathbb{P}(\hat{y} = 1|Y = 1, A)\Big).$$

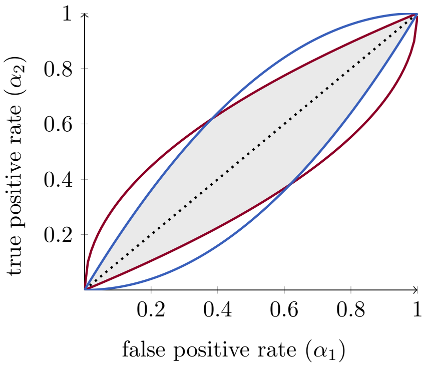


Figure 1: Achievable equal error rates (shaded). Two pairs of ROC curves form the boundaries of $S(L)$ and $S(H)$. Points in the intersection $S(L) \cap S(H)$ correspond to equal error rate classifiers.

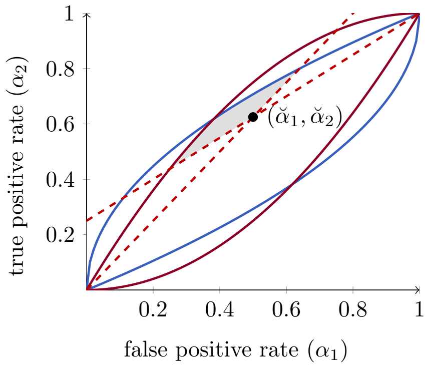


Figure 2: Achievable equal error rates from calibrated score at cutoff $\bar{p}$ (shaded). The restrictions (11) correspond to half-spaces above the red dashed lines.

We may now define the space of achievable error rates in $\mathbb{R}^2$. Let $\mathcal{H}$ be the set of all possibly random classifiers $h(X, A)$. The set of achievable error rates for group A is

$$S(A) = \{\alpha(\hat{y}, A) \,|\, \hat{y} = h(X, A), h \in \mathcal{H}\} \subseteq \mathbb{R}^2, \tag{7}$$

and the set of achievable rates for all classifiers satisfying equal error rates is given by $S(L) \cap S(H)$. To better understand this intersection, we characterize $S(A)$ in terms of Receiver Operator Characteristic (ROC) curves following Hardt et al. [11]. By definition, an ROC curve of a given score p traces the true and false positive rates associated with each possible cutoff rule $\mathbb{1}\{p \geq c\}$ for $c \in [0, 1]$. Therefore it contains all points $\alpha(\mathbb{1}\{p \geq c\}, A)$. With these tools in hand, we are ready to characterize the feasible space of rates $S(A)$ for group A.

Proposition 1. Let $p^* = p^*(X, A)$ be the Bayes optimal score satisfying $p^* = \mathbb{E}[Y|X, A]$, i.e., the best score given our data. Then the set of achievable rates $S(A)$ is exactly the convex hull of the union of the group-A ROC curve of the best score p^* and the group-A ROC curve of the worst score $1 - p^*$, i.e. the convex hull of

$$\left\{ \alpha(\mathbb{1}\{p^* \geq c\}, A) \,\middle|\, 0 \leq c \leq 1 \right\} \cup \left\{ (1, 1) - \alpha(\mathbb{1}\{p^* \geq c\}, A) \,\middle|\, 0 \leq c \leq 1 \right\}$$

Figure 1 illustrates typical examples of $S(L)$, $S(H)$, and the intersection $S(L) \cap S(H)$ which represents the rates achievable by equal error rate classifiers.

2.3.2 Classifiers Compatible with Calibration

We now put aside the equal error rate constraint and concentrate on identifying the entire set of classifiers that are implementable with the cutoff $\bar{p}$ applied to some calibrated scores $\hat{p}$. The set is characterized by the following proposition.

Proposition 2. A classifier $\hat{y}$ can be written as $\hat{y} = \mathbb{1}\{\hat{p} \geq \bar{p}\}$ for some calibrated $\hat{p}$ if and only if its group-specific positive predictive values exceed $\bar{p}$, and its group-specific negative predictive values exceed $1 - \bar{p}$. In particular, for $A \in \{L, H\}$,

$$\mathbb{P}(Y = 1|\hat{y} = 1, A) \geq \bar{p}, \quad \mathbb{P}(Y = 0|\hat{y} = 0, A) > 1 - \bar{p}. \tag{8}$$

Proof. Suppose that $\hat{y} = \mathbb{1}\{\hat{p} \geq \bar{p}\}$ where $\hat{p}$ is calibrated. Then $\hat{y}$ must satisfy the inequalities

$$\begin{aligned} \mathbb{P}(Y = 1|\hat{y} = 1, A) &= \mathbb{E}[Y|\hat{p} \geq \bar{p}, A] \\ &= \mathbb{E}[\hat{p}|\hat{p} \geq \bar{p}, A] \geq \bar{p}, \end{aligned} \tag{9}$$

$$\mathbb{P}(Y = 1|\hat{y} = 0, A) = \mathbb{E}[\hat{p}|\hat{p} < \bar{p}, A] < \bar{p}. \tag{10}$$

Therefore, if $\hat{y}$ is based on a calibrated score $\hat{p}$ at cutoff $\bar{p}$, then it is necessary for the group-specific positive and negative predictive values to exceed $\bar{p}$ and $(1 - \bar{p})$, respectively.

Conversely, given *any* classifier $\hat{y}$ that satisfies the inequalities (9) and (10), we can always put

$$\hat{p}(\hat{y}, A) = \mathbb{P}(Y = 1|\hat{y}, A)$$

to obtain a calibrated score that takes just two possible values per group with the cutoff $\bar{p}$ guaranteed to be between them. This choice of $\hat{p}$ thus satisfies $\hat{y} = \mathbb{1}\{\hat{p} \geq \bar{p}\}$ by construction. □

As we will see in the following subsection, this result lays the foundation for the necessary and sufficient conditions for the satisfiability of our fairness criteria.

2.3.3 The Feasibility Region

Proposition 2 demonstrates that the following are equivalent:

(i) There exists a calibrated score $\hat{p}$ such that $\hat{y} = \mathbb{1}\{\hat{p} \geq \bar{p}\}$ satisfies equal error rates.

(ii) There exists a classifier $\hat{y}$ satisfying equal error rates and (8).

In practice, we propose checking (ii) to identify whether (i) holds. To do so, we use Bayes' rule to write (8) as group-specific restrictions on true and false positive rates so that we can consider them in the same space as the equal error rate constraints given by Hardt et al. [11]. The following theorem and the accompanying Figure 2 indicates that each restriction (8) corresponds to a half-space in $\mathbb{R}^2$, and that the fairness feasibility region corresponds to the intersection of those half-spaces with each other and with the equal error rates region $S(L) \cap S(H)$.

Theorem 2. Let $\beta_A = \mu_A/(1 - \mu_A)$ denote the group-specific odds ratios, with $\beta_L < \beta_H$. Then our fairness criteria are simultaneously satisfiable at cutoff $\bar{p}$ if and only if there exists $(\alpha_1, \alpha_1) \in S(L) \cap S(H)$ satisfying the two inequalities

$$\frac{\alpha_2}{\alpha_1} \geq \frac{\bar{p}}{\beta_L(1-\bar{p})}, \quad \frac{(1-\alpha_1)}{(1-\alpha_2)} > \frac{\beta_H(1-\bar{p})}{\bar{p}} \tag{11}$$

We next provide easily checkable necessary and sufficient conditions for fairness feasibility

Corollary 1. Let $(\breve{\alpha}_1, \breve{\alpha}_2)$ denote the point at which the inequalities (11) hold with equality. Our fairness criteria are simultaneously satisfiable at cutoff $\bar{p}$ if and only if any of the following holds: $\breve{\alpha}_1 \leq 0$, $\breve{\alpha}_1 \geq 1$, or both groups' ROC curves corresponding to p^* lie above $(\breve{\alpha}_1, \breve{\alpha}_2)$. Note that $(\breve{\alpha}_1, \breve{\alpha}_2)$ are fixed by the group base rates and decision-maker's cutoff $\bar{p}$,

$$\breve{\alpha}_1 = \frac{\beta_L}{(\beta_H - \beta_L)}\left(\frac{\beta_H - (1+\beta_H)\,\bar{p}}{\bar{p}}\right), \quad \breve{\alpha}_2 = \frac{1}{(\beta_H - \beta_L)}\left(\frac{\beta_H\,(1-\bar{p}) - \bar{p}}{1-\bar{p}}\right). \tag{12}$$

Remark. The ROC curves correspond to the Bayes optimal score $p^* = \mathbb{E}[Y|X, A]$, which needs to be estimated in practice.

We close the section with a discussion of how data contributes to fairness feasibility, as illustrated by Theorem 2 and Figure 2. Note that the intersection of the half-spaces defined in (11) are fixed by given parameters: β_L and β_H through the population base rates, and $\bar{p}$ through the decision-maker's relative loss k from false positives. If the lines corresponding to those half-spaces intersect between 0 and 1, then what determines fairness feasibility is the height of the ROC curves.

Higher ROC curves correspond to more accurate predictions, which can be achieved by including more informative features X. This expands the region $S(H) \cap S(L)$ and thus always weakens the constraints dictating whether equal error rates and calibration are compatible in a given setting. Therefore, increasing the quality of data that an algorithm can access promotes our notions of fairness, whereas removing data compromises them.

3 A Loss-Minimizing Algorithm

After checking that our fairness criteria are feasible in a given setting, a natural next step is to search for the optimal fair solution, i.e. to identify the most informative score $\hat{p}$ that minimizes the decision-maker's loss subject to our fairness constraints. Our strategy is to first estimate the most accurate score $p^* = \mathbb{E}[Y|X, A]$ without regard to fairness, and then to use the estimate in two separate stages. First, identify the error rates that minimize loss subject to the fairness conditions (Section 3.1). Second, identify the most informative calibrated $\hat{p}$ that gives rise to those error rates at the decision-maker's cutoff $\bar{p}$ (Section 3.2). Generalizations of the algorithm are developed in the appendix and listed at the end of this section.

3.1 Error Rate Optimization

The first stage of the algorithm identifies the achievable error rates that minimize loss. We consider two formulations. The first minimizes decision-maker loss subject to full enforcement of our fairness criteria. The second, meanwhile, flexibly accommodates users who seek partial enforcement of the fairness criteria.

3.1.1 Basic Formulation

Let R denote the set of points (α_1, α_2) in the feasible region, i.e. the points in $S(H) \cap S(L)$ that satisfy (11). Note that R is necessarily convex, as it is the intersection of four convex regions: $S(H)$, $S(L)$, and the half-spaces satisfying (11). Moreover, according to the decision-maker's loss function, the classifier corresponding to (α_1, α_2) obtains expected loss $k\alpha_1(1 - \mathbb{E}[Y]) + (1 - \alpha_2)\mathbb{E}[Y]$. Thus, the optimal error rates minimize

$$\ell(\alpha_1, \alpha_2) \equiv k\alpha_1(1 - \mathbb{E}[Y]) + (1 - \alpha_2)\mathbb{E}[Y]. \tag{13}$$

over $(\alpha_1, \alpha_2) \in R$. The optimal $z^* = (\alpha_1^*, \alpha_2^*)$ selected will be on the top-left frontier of the feasible region in Figure 2, with the precise point on the frontier determined by the decision-maker's relative preference k over false positive and false negative classifications.

3.1.2 Flexible Formulation

An alternative formulation featuring a weighted error-rate penalty can accommodate multiple cases encountered in practice. Users can flexibly trade off fairness and accuracy objectives, minimize error disparities rather than eliminate them when the feasible region R is empty, and enforce just one error constraint as in the "equality of opportunity" criterion of [11].

First we define a broader domain for the algorithm to search over in place of R. It contains all the error rates implementable by a calibrated score at the decision-maker's cutoff, without regard to equal error rates. The domain is $R(H) \times R(L)$ where

$$R(A) = \left\{ (\alpha_1, \alpha_2) \in S(A) \,\middle|\, \frac{1 - \alpha_2}{1 - \alpha_1} < \frac{\bar{p}/\beta_A}{(1 - \bar{p})} \leq \frac{\alpha_2}{\alpha_1} \right\}.$$

We also replace the loss function (13). The generalized loss function takes group-specific error rates $z_A = (\alpha_{1A}, \alpha_{2A})$ and outputs a weighted sum of the decision-maker's expected loss from those rates and the disparities across them. The loss is

$$\gamma\ell(z_L) + (1 - \gamma)\ell(z_H) + (z_L - z_H)'\Lambda(z_L - z_H) \tag{14}$$

where $\ell(z_A)$ is the decision-maker's expected loss $k\alpha_{1A}(1 - \mathbb{E}[Y|A]) + (1 - \alpha_{2A})\mathbb{E}[Y|A]$ and γ is the fraction of individuals in group L. Meanwhile, Λ is a positive semidefinite matrix that provides the flexibility of varying the enforcement of error rate minimization. For example, taking $\Lambda = \lambda I$ for arbitrarily large λ recovers the equal error rate solution when the feasible region R is nonempty, and otherwise outputs the solution that minimizes error rate disparities. Meanwhile a small choice of λ places relatively more weight on accuracy.

Finally, Λ could be chosen so that differences in the true and false positive rates are weighted differently. For example, we can achieve equal true positive rates [11] by letting $\Lambda(2, 2)$ be large and assigning 0 to all other elements in Λ.

As a result of the flexible procedure, group-specific error rates z_L^* and z_H^* are chosen to minimize the generalized loss function (14). In the second stage of the algorithm, discussed next, users can identify a calibrated score that yields those target rates at the decision-maker's classification cutoff.

3.2 Risk Score Optimization

Once a feasible set of error rates is chosen, the decision-maker's expected loss is determined. However, multiple choices of calibrated scores may be compatible with those target rates at the cutoff $\bar{p}$, and we expect that in practice, decision-makers would prefer using the most informative scores. This section thus describes a method to recover the MSE-minimizing score $\hat{p}$ that implements the target rates z_L^* and z_H^* by solving a constrained optimal transport problem [21].

We base the method on the observation that the best fair $\hat{p}$ is recoverable through post-processing the Bayes optimal score $p^* = \mathbb{E}[Y|X, A]$. Hardt et al. [11] also justify post-processing in their development of an algorithm achieving equal error rate classifiers, and we adapt their argument to our setting in the appendix. We also discuss how our procedure can be thought of as finding the smallest *mean-preserving contraction* of p^* that yields the targeted error rates. Readers may note that this involves some randomization of scores. We explore the effects of the randomization empirically in our appendix, and meanwhile highlight that our algorithm's accuracy objective limits the extent to which scores change.

Our method defines one linear program per group A and seeks the most informative $\hat{p}_A$ such that

$$\alpha(\mathbb{1}\{\hat{p}_A \geq \bar{p}\}, A) = z_A^* = (\alpha_{1A}^*, \alpha_{2A}^*).$$

For the remainder of the section, we simplify notation by suppressing A subscripts and note that the procedure is performed once for each group $A \in \{H, L\}$.

Our approach will involve a transformation kernel that maps the distribution of the most accurate p^* to the distribution of our post-processed $\hat{p}$. We assume for simplicity (with justification in the appendix) that p^* is discrete and takes N ordered values $p = (p_1, p_2, \dots, p_N)$, each with probability mass given by $s = (s_1, s_2, \dots, s_N)$ where $\sum_i s_i = 1$. Furthermore, we will denote the post-processed $\hat{p}$ as taking those same discrete values p but with different probability masses that we seek to optimize, $f = (f_1, f_2, \dots, f_N)$.

We call T the matrix that maps probability masses from the discrete distribution of p^* to that of $\hat{p}$. In particular, with probability T_{ij}, the kernel will map an individual with score p_i to the output score p_j. Therefore, the probability distribution of $\hat{p}$ will be determined by

$$T's = f. \tag{15}$$

In order to produce probability distributions, T must be right-stochastic: elements must take values between 0 and 1, and each row should sum to 1.

$$0 \leq T_{ij} \leq 1 \text{ and } \sum_{k=1}^{N} T_{ik} = 1 \quad \forall\, i, j \in \{1, \dots N\}. \tag{16}$$

According to our fairness criteria, we further constrain T. To ensure that $\hat{p}$ will be calibrated, we need the outcome of individuals assigned score f_i to satisfy $Y = 1$ with probability p_i. Assuming that p^* is itself calibrated (relaxed in the appendix), this reduces to

$$\sum_{i=1}^{N} T_{ij} p_i s_i = p_j f_j \quad \forall\, j \in \{1, \dots, N\}. \tag{17}$$

The targeted false- and true-positive rates (α_1^*, α_2^*) derived in Section 3.1 similarly require:

$$\sum_{j=1}^{N}\sum_{i=1}^{N} T_{ij} p_i s_i \left(\mathbb{1}\{p_j \geq \bar{p}\} - \alpha_2^*\right) = 0, \quad \sum_{j=1}^{N}\sum_{i=1}^{N} T_{ij}(1 - p_i) s_i \left(\mathbb{1}\{p_j \geq \bar{p}\} - \alpha_1^*\right) = 0. \tag{18}$$

Finally, we formulate an objective. Note that the mean-squared error of $\hat{p}$ satisfies the bias-variance decomposition

$$\mathbb{E}[(\hat{p} - Y)^2] = \mathbb{E}[(\hat{p} - \mathbb{E}[Y|X, A])^2] + \mathbb{E}[(Y - \mathbb{E}[Y|X, A])^2],$$

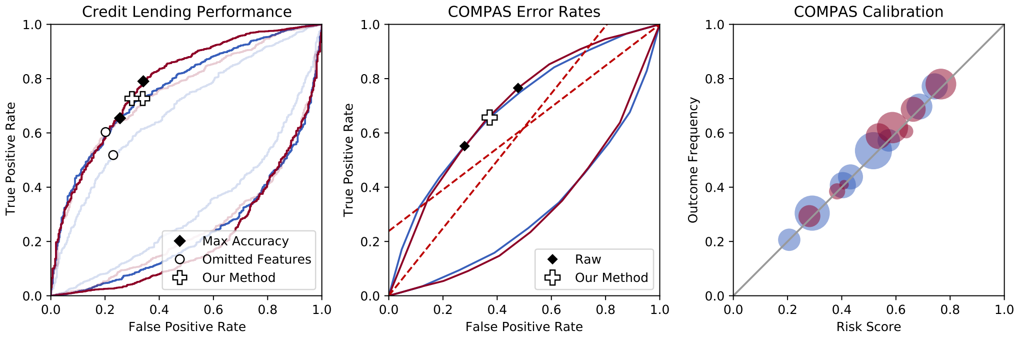
and thus the $\hat{p}$ that minimizes the left hand side is obtained by minimizing the first term on the right hand side. In particular, if the input score p^* is $\mathbb{E}[Y|X, A]$, then the post-processed score that minimizes mean-squared error will also minimize

$$\mathbb{E}[(\hat{p} - p^*)^2] = \sum_{i=1}^{N}\sum_{j=1}^{N} T_{ij}(p_i - p_j)^2 s_i. \tag{19}$$

Furthermore, even if p^* is not exactly equal to $\mathbb{E}[Y|X, A]$, the triangle inequality in $L^2(\mathbb{P})$ implies

$$\mathbb{E}[(\hat{p} - Y)^2]^{\frac{1}{2}} \leq \mathbb{E}[(p^* - Y)^2]^{\frac{1}{2}} + \mathbb{E}[(\hat{p} - p^*)^2]^{\frac{1}{2}}.$$

Thus, by minimizing the objective (19) we can effectively control the additional error due to post-processing. Combining this with the above constraints yields a straightforward linear program.



(a)

(b)

(c)

Figure 3: Evaluating algorithm performance. In each figure, maroon represents the high-mean group while blue represents the low-mean group. Panel (a) corresponds to credit lending, illustrating the empirical ROC curves from the rich feature set (opaque) and the limited feature set (translucent). Panels (b) and (c) cover the criminal justice application, showing respectively that we can eliminate error rate disparities and maintain score calibration. Note that we define a true positive classification as correctly identifying someone who would not reoffend.

3.3 Available Generalizations

Interested readers are invited to reference the appendix for further discussion of this algorithm and additional generalizations. The procedure is shown to handle cases where the decision-maker's cutoff $\bar{p}$ is estimated with error and where there are more than two groups.

4 Empirical Results

Let us take our procedure to data. In the first application, we design a risk score to aid a lender's classification task to authorize loans, showing that it outperforms a common alternative strategy based on the omission of sensitive features. Afterwards, we build on the criminal justice algorithm COMPAS to demonstrate that risk assessments can be designed to output both calibrated risk scores as well as equal error rate binary risk summaries. Detailed discussion of each example is included in our appendix for interested readers.

4.1 Predicting loan repayment

We first present an example of designing a risk score to inform a credit lenders approvals of loan applicants. Our predictions are constructed using the Survey of Income and Program Participation (SIPP), a nationally-representative survey of the civilian population spanning multiple years [23]. We select as our outcome the ability to pay rent, mortgage, and utilities in 2016, and suppose that we are tasked with predicting that outcome using survey responses from two years prior. Furthermore we assume that the lender regards missed payments as highly costly and accordingly authorizes loans only to those with scores greater than $\bar{p} \approx .909$, corresponding to $k = 10$. Our goal is to design scores that treat high-educated and low-educated applicants fairly, so that creditworthy individuals will have the same probability of being granted a loan regardless of their education.

The full dataset contains over 1,800 features spanning detailed financial variables (including work history, assets, and debts), as well as sensitive features (including demographic information). We apply our algorithm to the full feature set and derive calibrated scores that yield equal true positive rates at the lender's cutoff. Then, we compare its performance to an accuracy-maximizing algorithm that uses the commonly practiced fairness strategy of omitting all sensitive features from the training data. The results are summarized numerically in Table 1 and graphically in Figure 3a. Our algorithm simultaneously achieves lower loss for the lender as well as higher probabilities that creditworthy applicants of both education groups are granted loans.

Table 1: Application to credit lending. Row [1] is based on raw scores. Row [2] summarizes the classifier that minimizes lender loss subject to equal true positive rates. Our algorithm summarized in row [3] produces a calibrated score corresponding to equal true positive rate classifications; since it retrieves the same error rates as row [2], we see there is no added loss from enforcing score calibration. Row [4] summarizes the scores from the alternative procedure that omits sensitive features, displaying greater loss for the lender, lower true positive rates for both groups, and substantial error disparities across groups.

	Algorithmic Target	Lender Loss	TPR (H/L)	FPR (H/L)	Score MSE
	Trained on all features				
[1]	Accuracy Maximizing	.517	(.795/.661)	(.341/.255)	.072
[2]	Eq. TPR Only	.532	(.727/.727)	(.299/.339)	N/A
[3]	*Eq. TPR + Calibration*	.532	(.727/.727)	(.299/.339)	.073
	Trained on limited features				
[4]	Accuracy Maximizing	.591	(.603/.518)	(.202/.230)	.077

4.2 Predicting criminal recidivism

In a second application, our procedure can design risk assessments that output both calibrated scores as well as binary high or low risk summaries satisfying equal error rates. To illustrate, we consider the COMPAS criminal justice algorithm. In 2016, *ProPublica* reported that the algorithms risk summaries displayed substantial error imbalances across race, although the algorithms scores satisfied predictive parity overall [2, 1].

To check whether we can correct COMPAS error imbalances without sacrificing score calibration, we applied our post-processing technique to Broward County risk scores made public by *ProPublica* [17]. Motivated by *ProPublica*'s influential analysis [18], we define the outcome as the measure of recidivism within two years, and suppose that all defendants with scores 1 through 4 are flagged to a judge as "low" risk, while those from 5 to 10 are flagged as "high." We derived the associated ROC curves of the scores, as seen in Figure 3b. Our procedure eliminates the racial error disparities of the associated risk classifications. As seen in Figure 3c, it also preserves calibration.

5 Conclusion

In settings from hospitals to courtrooms, decision-makers stand to benefit from algorithmic predictions. This paper studies fair prediction in the widespread setting where a risk score is constructed to aid their classification tasks. We prove that it is possible to construct calibrated scores that lead to equal error rate classifications at group-blind cutoffs. We characterize exactly when a solution is possible and propose an algorithm that produces the most informative score satisfying the fairness criteria and minimizing the decision-makers errors. Finally, we emphasize the importance of data richness to fairness. Compared to a commonly practiced strategy of omitting sensitive data, our algorithm can produce scores that enhance both efficiency and equity.

6 Thoughts on Broader Impact

Today's algorithmic risk scores form the basis for decisions that directly impact people. Risk scores can determine whether a child welfare official opens an abuse investigation into a family, whether a physician enrolls a patient into a treatment program, whether a judge chooses the grant bail to a defendant, and whether a lender authorizes a mortgage for a family seeking to buy a home. In this paper, we have sought to expand the range of tools available to the developers of these modern risk assessments.

There are two sets of scenarios where our work can be applied, depending on whether it is a risk assessment or a third-party decision-maker that is responsible for deriving classifications. The first case accommodates a risk assessment tool similar to COMPAS that displays to its user both a risk score as well as a corresponding high or low risk summary for each individual. Our procedure could be used to produce calibrated scores that, at a designated cutoff chosen by the developer, correspond to binary summaries satisfying equal error rates.

In the second set of possible applications, our procedure can produce calibrated risk scores for a decision-maker tasked with assigning binary treatments, such as authorizing or denying loans [15, 12]. For the procedure to successfully induce equal error rate outcomes, knowledge is required about how the decision-maker weighs the costs of false positive and negative classifications. That is, collaboration with the decision-maker during the design process is crucial.

When considering the social implications of this work, we emphasize that the two fairness criteria that we have focused on here do not encompass all notions of fairness. Tradeoffs remain between these criteria and others. For example, enforcing equal error rates requires that the classifications positive and negative predictive values will be unequal across groups, meaning that one groups scores would carry greater signal to the decision-maker than the others [5]. Perhaps more importantly, equal error rate classifications will generically require changes to the Bayes optimal classifications that favor certain groups, and enforcing calibration does not diminish this requirement. Relatedly, implementing equal error rates across one group identifier can sometimes cause imbalances across other group identifiers [6].

We believe that every prediction setting warrants individualized fairness assessments and a tailored approach. The choice of how to prioritize fairness conditions is ultimately up to human beings. We hope that by clarifying the precise relationship between two influential criteria, we can facilitate these decisions, and that in settings where calibration and equal error rates are considered essential, our algorithm can help yield accurate predictions and fairer outcomes.

Acknowledgments

Many thanks to Anna Mikusheva, Iván Werning, and Ben Deaner for their valuable advice. We're also deeply grateful for the support of David Autor, Lou Crandall, Pari Sastry, Tom Brennan, Jim Poterba, Rachael Meager, and Frank Schilbach with whom we have had energizing and productive conversations. Thank you to Pooya Molavi, Adam Fisch, and Yonadav Shavit for commenting on the manuscript at its advanced stages. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1122374.

References

[1] Julia Angwin and Jeff Larson. Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say. https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say, December 2016.

[2] Julia Angwin and Jeff Larson. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, May 2016.

[3] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104 (3):671–732, 2016. ISSN 00081221.

[4] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, page 0049124118782533, 2018.

[5] Alexandra Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163, June 2017. ISSN 2167-6461. doi: 10.1089/big.2016.0047.

[6] Sam Corbett-Davies and Sharad Goel. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv:1808.00023 [cs]*, August 2018.

[7] Sam Corbett-Davies, Emma Pierson, Avi Feller, and Sharad Goel. A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/, 2016.

[8] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining, KDD ’17, pages 797–806, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 978-1-4503-4887-4. doi: 10.1145/3097983.3098095.

[9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS ’12, pages 214–226, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 978-1-4503-1115-1. doi: 10.1145/2090236.2090255.

[10] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 119–133, New York, NY, USA, February 2018. PMLR.

[11] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of Opportunity in Supervised Learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3315–3323. Curran Associates, Inc., 2016.

[12] Sampath Kannan, Michael Kearns, Jamie Morgenstern, Mallesh Pai, Aaron Roth, Rakesh Vohra, and Zhiwei Steven Wu. Fairness incentives for myopic agents. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, EC ’17, pages 369–386, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 978-1-4503-4527-9. doi: 10.1145/3033274.3085154.

[13] Jon Kleinberg and Sendhil Mullainathan. Simplicity creates inequity: Implications for fairness, stereotypes, and interpretability. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC ’19, pages 807–808, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 978-1-4503-6792-9. doi: 10.1145/3328526.3329621.

[14] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv:1609.05807 [cs, stat]*, November 2016.

[15] 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, February 2018. ISSN 0033-5533. doi: 10.1093/qje/qjx032.

[16] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic Fairness. *AEA Papers and Proceedings*, 108:22–27, 2018. ISSN 2574-0768. doi: 10.1257/pandp.20181018.

[17] Jeff Larson. Data and analysis for ‘Machine bias’. June 2017.

[18] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[19] Marc Mezard and Andrea Montanari. *Information, Physics, and Computation*. Oxford University Press, Inc., USA, 2009. ISBN 0-19-857083-X.

[20] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D’Amour, and Kristian Lum. Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. *arXiv:1811.07867 [stat]*, July 2019.

[21] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[22] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On Fairness and Calibration. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5680–5689. Curran Associates, Inc., 2017.

[23] U.S. Census Bureau. Survey of income and program participation. https://www.census.gov/programs-surveys/sipp/data/datasets.html, 2014.

[24] Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2008.

[25] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1171–1180, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4913-0. doi: 10.1145/3038912.3052660.

Appendix for Section 2

Omitted proofs in 2.2

Addendum to Theorem 1

Addendum. We relax condition (iii) of Theorem 1 and replace it with the weaker condition that $\mathrm{Var}(Y|Z) > \epsilon$ almost surely. This will correspond to the assumption that Y cannot be perfectly predicted from any realization of Z. We will make use of the criterion that Borel random variables R and R' are independent conditional on Σ iff for all bounded, continuous f and g we have

$$\mathbb{E}[f(R)g(R')|\Sigma] = \mathbb{E}[f(R)|\Sigma]\mathbb{E}[g(R')|\Sigma].$$

Now suppose that (Z, A, Y) are known to satisfy Theorem 1 conditions (i) and (ii), and that $\mathrm{Var}(Y|Z) > 0$. Then let η be a $\mathrm{Ber}(\varepsilon)$ random variable independent of (Z, A, Y). We consider a variable A_η that takes value A with probability $1 - \varepsilon$ and otherwise flips the variable A with probability ε, that is,

$$A_\eta = A + \eta \pmod 2.$$

This gives us a triple (Z, A_η, Y) that satisfies $\mathbb{E}[A|Z] \in (0, 1)$ and $\mathbb{E}[Y|A, Z] \in (0, 1)$ almost surely by construction, corresponding to condition (iii) from the Theorem. We can also show that the triple satisfies the other two conditions. For instance, to show that condition (i) holds, let S be an arbitrary set such that $S \in \sigma(Z)$. We will use the fact that any $\sigma(Z)$-measurable random variable V and any random variable U satisfy $\mathbb{E}[\mathbb{E}[U|Z]V] = \mathbb{E}[UV]$. In particular, because $\mathbb{1}_{Z\in S}$ is $\sigma(Z)$-measurable.

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}[f(A_\eta)g(Y)|Z]\mathbb{1}_{Z\in S}\right] &= \mathbb{E}\left[f(A_\eta)g(Y)\mathbb{1}_{Z\in S}\right] \\
&= \mathbb{E}_{(A,Z,Y)}\left[\mathbb{E}_\eta[f(A_\eta)]g(Y)\mathbb{1}_{Z\in S}\right] \text{ by independence of } \eta \\
&= \mathbb{E}_{(A,Z,Y)}\left[\mathbb{E}[\mathbb{E}_\eta[f(A_\eta)]g(Y)|Z]\mathbb{1}_{Z\in S}\right] \text{ since } \mathbb{1}_{Z\in S} \text{ is } \sigma(Z)\text{-measurable} \\
&= \mathbb{E}_{(A,Z,Y)}\left[\mathbb{E}[\mathbb{E}_\eta[f(A_\eta)]|Z]\mathbb{E}[g(Y)|Z]\mathbb{1}_{Z\in S}\right] \text{ by assumption that } (Y \perp\!\!\!\perp A) \mid Z \\
&= \mathbb{E}_{(A,Z,Y)}\left[\mathbb{E}_\eta[f(A_\eta)]\mathbb{E}[g(Y)|Z]\mathbb{1}_{Z\in S}\right] \text{ by independence of } \eta \\
&= \mathbb{E}\left[f(A_\eta)\mathbb{E}[g(Y)|Z]\mathbb{1}_{Z\in S}\right] \\
&= \mathbb{E}\left[\mathbb{E}[f(A_\eta)|Z]\mathbb{E}[g(Y)|Z]\mathbb{1}_{Z\in S}\right] \text{ since } \mathbb{E}[g(Y)|Z]\mathbb{1}_{Z\in S} \text{ is } \sigma(Z)\text{-measurable.}
\end{aligned}$$

Because S was arbitrary and both $\mathbb{E}[f(A_\eta)|Z]\mathbb{E}[g(Y)|Z]$ and $\mathbb{E}[f(A_\eta)g(Y)|Z]$ are $\sigma(Z)$-measurable, we can conclude that $\mathbb{E}[f(A_\eta)|Z]\mathbb{E}[g(Y)|Z] = \mathbb{E}[f(A_\eta)g(Y)|Z]$ almost surely so (i) is satisfied. A very similar argument shows that (ii) holds. Therefore, by Theorem 1, A_η is independent of (Z, Y). Then given arbitrary bounded and continuous functions f and g,

$$\mathbb{E}[f(A_\eta)g(Z, Y)] = \mathbb{E}[f(A_\eta)]\mathbb{E}[g(Z, Y)].$$

Using the fact that $A_\eta \to A$ as $\eta \downarrow 0$ in $L^2(\mathbb{P})$, and that $h \mapsto \mathbb{E}[h]$ and $(h, h') \mapsto \mathbb{E}[hh']$ are continuous in $L^2(\mathbb{P})$, we conclude by continuity that

$$\mathbb{E}[f(A)g(Z, Y)] = \mathbb{E}[f(A)]\mathbb{E}[g(Z, Y)].$$

Since f and g were arbitrary, we have in fact shown that A is independent of (Z, Y), as wanted.

Thus, we have succeeded in proving the following refinement: under Theorem 1 assumptions (i) and (ii), if Y cannot be perfectly predicted from any realization of Z, then the random variables A and (Y, Z) must be independent.

Since assumptions (i) and (ii) continue to hold if we condition on $Z \in S$ for any S, we can say further that if Theorem 1 conditions (i) and (ii) hold and P is the set of values of Z from which perfect prediction is not possible, i.e. $\mathrm{Var}(Y|Z) > 0$ then A and Y are independent conditionally on $Z \in P$.

□

Omitted proofs in 2.3.1

Proof of Proposition 1

Proof. First we can prove a lemma stating that $S(A)$ is convex. To see this, let ξ be an independent $\mathrm{Ber}(\lambda)$ random variable. Then, by iterating expectations, one sees that

$$\alpha(\hat{y} + \xi(\hat{z} - \hat{y}), A) = \lambda\alpha(\hat{z}, A) + (1 - \lambda)\alpha(\hat{y}, A).$$

Using this convexity, we can prove the proposition. Note that the points $\alpha(\mathbb{1}\{p^* \geq c\}, A)$ that make up the group-A ROC curve of p^* describe the error rates achieved by all cutoff classifiers based on p^*, and so they are in $S(A)$. Meanwhile, since

$$\alpha(1 - \hat{y}, A) = (1, 1) - \alpha(\hat{y}, A),$$

the points $(1, 1) - \alpha(\mathbb{1}\{p^* \geq c\}, A)$ must also be in $S(A)$. This corresponds to the group-A ROC curve of the scores $1 - p^*$. Any point in the convex hull of these two ROC curves can be achieved by randomization as in the aforementioned lemma. For further details and intuition, see Section 4 in Hardt *et al.* (2016). Note that Hardt *et al.* choose not to illustrate the feasible region below the main diagonal as it corresponds to classifiers that are worse than random.

To show that *all* attainable error rates belong to this set, we use the convexity of $S(A)$ to note that the support points of $S(A)$ correspond to all classifiers that yield extrema of $\gamma_1\alpha_1(\hat{y}, A) + \gamma_2\alpha_2(\hat{y}, A)$ where (γ_1, γ_2) are arbitrary weights. To describe these support points tractably, we can use the result derived later in the appendix (Proposition A.3) that shows that optimal classifications can be chosen to depend on only p^* and A, where $p^* = \mathbb{E}[Y|X, A]$. Thus the extrema of $\gamma \cdot \alpha(\hat{y}, A)$ are achieved by cutoff rules $f(p^*, A) = \mathbb{1}\{p^* \geq c\}$ and $f(p^*, A) = \mathbb{1}\{p^* < c\}$, giving support points

$$\alpha(\mathbb{1}\{p^* \geq c\}, A)$$

and

$$\alpha(\mathbb{1}\{p^* < c\}, A) = (1, 1) - \mathbb{1}\{p^* \geq c\}, A)$$

which as we have shown are all contained in $S(A)$. Finally, we use the fact that a convex set containing all of its support points is equal to the convex hull of its support points. □

Omitted proofs in 2.3.3

Extension of Proposition 2

Proof. Suppose that (i) holds and call $\hat{p}_f$ the fair score for which $\hat{y} = \mathbb{1}\{\hat{p}_f \geq \bar{p}\}$ satisfies equal error rates. Then since $\hat{p}_f$ is calibrated,

$$\begin{aligned}\mathbb{P}(Y = 1|\hat{y} = 1, A) &= \mathbb{E}[Y|\hat{p}_f \geq \bar{p}, A] = \\ &= \mathbb{E}[\hat{p}_f|\hat{p}_f \geq \bar{p}, A] \geq \bar{p}\end{aligned}$$

and similarly,

$$\mathbb{P}(Y = 1|\hat{y} = 0, A) < \bar{p}$$

So in addition to satisfying equal error rates, $\hat{y}$ satisfies (9) and (10), which are equivalent to the two conditions in (8). Thus (ii) is a necessary condition for fairness.

Now we show the converse; (ii) is also sufficient for fairness. Suppose that (ii) holds and let $\hat{y}_f$ be a classifier satisfying equal error rates and (8). Choose $\hat{p}(\hat{y}_f, A) = \mathbb{P}(Y = 1|\hat{y}_f, A)$. These scores are calibrated by construction. Also, since they satisfy $\hat{p}(\hat{y}_f = 0, A) < \bar{p}$ and $\hat{p}(\hat{y}_f = 1, A) \geq \bar{p}$, they exactly implement the classifier $\hat{y}_f$ at the cutoff $\bar{p}$. □

Proof of Theorem 2

Proof. Building on the above extension of Proposition 2, it is enough for us to show that the existence of the point $(\alpha_1, \alpha_2) \in S(L) \cap S(H)$ satisfying (11) is equivalent to the following: There exists a classifier $\hat{y}$ satisfying equal error rates and (8).

First note that $S(L) \cap S(H)$ is nonempty, since for example $(0, 0)$ and $(1, 1)$ are points in both $S(L)$ and $S(H)$. So we can consider some arbitrary (α_1, α_2) that is in $S(L) \cap S(H)$ and is therefore implementable by an equal error rate classifier that we call $\hat{y}_e$. We need to show that $\hat{y}_e$ satisfying the conditions in (8) $\forall A$ is equivalent to its corresponding true and false positive rates $(\alpha_1(\hat{y}_e, A), \alpha_2(\hat{y}_e, A))$ satisfying (11) $\forall A$.

Recall that the PPV condition in (8) required

$$\mathbb{P}(Y = 1|\hat{y}_e = 1, A) \geq \bar{p}.$$

Applying Bayes' rule to the inequality, we have

$$\begin{aligned}\mathbb{P}(Y=1|\hat{y}_e=1,A) &= \frac{\mathbb{P}(\hat{y}_e=1|Y=1,A)\mathbb{P}(Y=1|A)}{\mathbb{P}(\hat{y}_e=1|A)} \\ &= \frac{\alpha_2(\hat{y}_e,A)\mu_A}{\alpha_2(\hat{y}_e,A)\mu_A+\alpha_1(\hat{y}_e,A)(1-\mu_A)} \\ &\geq \bar{p}\end{aligned}$$

After algebraic manipulation, the restriction can be written

$$\frac{\alpha_2(\hat{y}_e,A)}{\alpha_1(\hat{y}_e,A)} \geq \frac{\bar{p}(1-\mu_A)}{(1-\bar{p})\mu_A} = \frac{\bar{p}}{(1-\bar{p})\beta_A}$$

where $\beta_A \equiv {}^{\mu_A}/_{(1-\mu_A)}$. Therefore $(\alpha_1(\hat{y}_e,A), \alpha_2(\hat{y}_e,A))$ must satisfy the following for both $A=L$ and $A=H$

$$\frac{\alpha_2(\hat{y}_e,A)}{\alpha_1(\hat{y}_e,A)} \geq \frac{\bar{p}}{(1-\bar{p})\beta_A}$$

Since $\beta_L < \beta_H$, the condition is more restrictive when $A=L$, giving the first condition in (11). We next similarly transform the NPV condition in (8), recalling it requires

$$\mathbb{P}(Y=0|\hat{y}=0,A) > 1-\bar{p}$$

and by Bayes' rule,

$$\begin{aligned}\mathbb{P}(Y=0|\hat{y}=0,A) &= \frac{\mathbb{P}(\hat{y}=0|Y=0,A)\mathbb{P}(Y=0|A)}{\mathbb{P}(\hat{y}=0|A)} \\ &= \frac{(1-\alpha_1(\hat{y},A))(1-\mu_A)}{(1-\alpha_1(\hat{y},A))(1-\mu_A)+(1-\alpha_2(\hat{y},A))\mu_A} \\ &> 1-\bar{p}\end{aligned}$$

After algebraic manipulation, this becomes $\forall A$

$$\frac{(1-\alpha_1(\hat{y},A))}{(1-\alpha_2(\hat{y},A))} > \frac{(1-\bar{p})\beta_A}{\bar{p}}$$

Since $\beta_H > \beta_L$, the most restrictive case is when $A=H$, giving the second condition in (11).

Note that special attention should be given to the corner solutions. At point $(0,0)$, the first condition in (11) becomes irrelevant and so the second condition in (11) is necessary and sufficient. Meanwhile at $(1,1)$, the second condition in (11) becomes irrelevant so the first condition in (11) is necessary and sufficient. □

Proof of Corollary 1

Proof. Let F and G denote the lines for which the inequalities (11) hold with equality. That is to say, $F, G \subset \mathbb{R}^2$ are given by

$$F = \left\{(\alpha_1,\alpha_2) \in \mathbb{R}^2 \,\middle|\, \frac{\alpha_2}{\alpha_1} = \frac{\bar{p}}{\beta_L(1-\bar{p})}\right\}$$
$$G = \left\{(\alpha_1,\alpha_2) \in \mathbb{R}^2 \,\middle|\, \frac{(1-\alpha_1)}{(1-\alpha_2)} = \frac{\beta_H(1-\bar{p})}{\bar{p}}\right\}$$

The lines intersect at $(\breve{\alpha}_1, \breve{\alpha}_2)$ given by (12). Our proof will rest on a few basic facts: $S(L) \cap S(H)$ is convex, F contains $(0,0)$, G contains $(1,1)$, and both lines have positive slope.

First we prove that if $\breve{\alpha}_1 \leq 0$, $\breve{\alpha}_1 \geq 1$, or both ROC curves lie above the intersection $(\breve{\alpha}_1, \breve{\alpha}_2)$, then there exists a point (α_1, α_2) satisfying the feasibility conditions in Theorem 2.

Case I: $0 < \breve{\alpha}_1 < 1$ and $(\breve{\alpha}_1, \breve{\alpha}_2)$ lies below both ROC curves. Note that increasing α_2 slackens both inequalities (11). Thus, if $0 < \breve{\alpha}_1 < 1$ and $(\breve{\alpha}_1, \breve{\alpha}_2)$ lies below both ROC curves, there then exists a point $(\breve{\alpha}_1, \alpha_2)$ with $\alpha_2 > \breve{\alpha}_2$ that lies on the minimum of the two ROC curves, hence in $S(H) \cap S(L)$, and moreover the inequalities (11) hold at $(\breve{\alpha}_1, \alpha_2)$. This is a feasible point.

Case II: $\breve{\alpha}_1 \leq 0$. On the other hand, if $\breve{\alpha}_1 \leq 0$, then in $(0,1) \times \mathbb{R}$ the line F lies strictly above G. Then the point $(0,0) \in S(L) \cap S(H) \cap F$ lies above G, meaning that the second condition in (11) holds and the point is feasible.

Case III: $\breve{\alpha}_1 \geq 1$. If $\breve{\alpha}_1 \geq 1$, then in $(0,1) \times \mathbb{R}$ the line G lies strictly above F. Then the point $(1,1) \in S(L) \cap S(H) \cap G$ lies above F, so the first condition in (11) holds and the point is feasible.

Finally, we prove the converse that if $0 < \breve{\alpha}_1 < 1$ and $\breve{\alpha}_2$ lies above at least one of the ROC curves, then the feasible region is empty. Let the intersection of $S(L) \cap S(H)$ with the half-space above F be denoted by I_F, and the intersection of $S(L) \cap S(H)$ with the half-space above G be denoted by I_G. We need to show that $I_F \cap I_G$ is empty. The argument follows from the convexity of $S(L) \cap S(H)$ and the fact that both F and G have positive slopes. In particular, due to the convexity of $S(L) \cap S(H)$, the positive slope of F, and the fact that $(0,0)$ is in F, we know the line F must intersect the boundary of $S(L) \cap S(H)$ strictly to the left of $\breve{\alpha}_1$. Meanwhile, G must intersect the boundary of $S(L) \cap S(H)$ strictly to the right of $\breve{\alpha}_1$. Thus the rightmost point of I_F lies strictly to the left of the leftmost point of I_G, and the intersection of $S(L) \cap S(H)$ with both half-spaces above F and G must be empty.

□

Appendix for Section 3

Justification for post-processing p^*

First we justify post-processing the Bayes optimal p^* to arrive at the optimal fair $\hat{p}$. To do so we adapt Proposition 5.2 from Hardt *et al.* (2016) to our setting and prove the following

Proposition A.3. For any source distribution over (Y, X, A) with Bayes optimal regressor given by $p^*(X, A) = \mathbb{E}[Y|X, A]$ and loss function ℓ, there exists a predictor $\hat{p}(p^*, A)$ such that

(i) $\hat{p}$ is an optimal predictor satisfying our fairness properties of calibration and equal error rates. That is, $\mathbb{E}[\ell(\mathbb{1}_{\hat{p} > \underline{p}}, Y)] \leq \mathbb{E}[\ell(\mathbb{1}_{\hat{g} > \underline{p}}, Y)]$ for any $\hat{g}$ that satisfies the properties.

(ii) $\hat{p}$ is derived from (p^*, A). In particular, it is a (possibly random) function of the random variables (p^*, A) alone, and is independent of X conditional on (p^*, A).

Proof. To start, first note that our fairness properties of calibration and equal error rates on a score p and classifications $\mathbb{1}\{p \geq \bar{p}\}$ are "oblivious." That is, they depend only on the joint distribution of (Y, A, p) given the known cutoff $\bar{p}$. We will show that for any arbitrary $\hat{g}$ that satisfies the fairness properties, we can construct a $\hat{p}$ that also satisfies fairness, yields the same expected loss, and is derived from (p^*, A).

Consider an arbitrary $\hat{g} = f(X, A)$ satisfying the fairness properties. We can define $\hat{p}(p^*, A)$ as follows: draw a vector X' independently from the conditional distribution of X given the realized values of p^* and A, and set $\hat{p} = f(X', A)$. Note this $\hat{p}$ satisfies (ii) by construction.

To show that this $\hat{p}$ satisfies the fairness properties and yields the same expected loss as $\hat{g}$, note that since Y is binary with conditional expectation equal to the Bayes optimal p^*, we know Y is independent of X conditional on p^*. Therefore (Y, p^*, X, A) and (Y, p^*, X', A) have the same joint distribution, and so must $(f(X, A), A, Y)$ and $(f(X', A), A, Y)$. Since the fairness properties are oblivious and depend only on these latter joint distributions, then we know that as long as $\hat{g}$ satisfies them then so will $\hat{p}$. Finally, we can deduce that $(Y, \hat{g})$ and $(Y, \hat{p})$ also have the same joint distribution, meaning that (i) is satisfied with equality. □

Mean-preserving contractions of calibrated scores

We observe that a calibrated score derived from another is a mean-preserving contraction. Since the Bayes optimal p^* that serves as input to our algorithm frequently satisfies calibration (see Liu *et al.* 2019), then our post-processing method can be viewed as finding its smallest mean preserving contraction that achieves equal error rates at the decision-maker's cutoff.

The relationship between calibrated scores related by post-processing is characterized by our proposition below.

Proposition A.4. Let p_A be any calibrated score of group A, i.e. satisfying $\mathbb{E}[Y|p_A] = p_A$ for members of A, and let $\hat{p}_A = f(p_A, \zeta)$ be a score post-processed from p_A that is also calibrated, where ζ is independent of Y conditional on p_A. Then, $\hat{p}_A$ is a mean-preserving contraction of p_A, with $p_A = \hat{p}_A + Z$ and $\mathbb{E}[Z|\hat{p}_A] = 0$. Conversely, any $\tilde{p}_A$ that satisfies $p_A = \tilde{p}_A + Z$ with $\mathbb{E}[Z|\tilde{p}_A] = 0$ is calibrated.

Proof. We first show that $\hat{p}_A$ is a mean-preserving contraction of p_A. To start, note that the post-processed $\hat{p}_A$ is assumed to be calibrated, so $\mathbb{E}[Y|\hat{p}_A] = \hat{p}_A$. Moreover, since $\hat{p}_A = f(p_A, \zeta)$, we have $\sigma(\hat{p}_A) \subseteq \sigma(p_A, \zeta)$. Therefore by the tower property of conditional expectation,

$$\begin{aligned}\hat{p}_A &= \mathbb{E}[Y|\hat{p}_A] \\ &= \mathbb{E}[\mathbb{E}[Y|p_A, \zeta]|\hat{p}_A] \\ &= \mathbb{E}[\mathbb{E}[Y|p_A]|\hat{p}_A] \text{ by conditional independence of } \zeta \\ &= \mathbb{E}[p_A|\hat{p}_A] \text{ by calibration of } p_A\end{aligned}$$

Then $p_A = p_A + (\hat{p}_A - \mathbb{E}[p_A|\hat{p}_A]) = \hat{p}_A + (p_A - \mathbb{E}[p_A|\hat{p}_A])$ where the second term is by construction mean independent of $\hat{p}_A$, so $\hat{p}_A$ is a mean-preserving contraction of p_A.

Now we show that if the score $\tilde{p}_A$ is a mean-preserving contraction of p_A such that $p_A = \tilde{p}_A + Z$ for some Z satisfying $\mathbb{E}(Z|\tilde{p}_A) = 0$, then $\tilde{p}_A$ is calibrated. Observe that

$$\begin{aligned}\mathbb{E}[p_A|\tilde{p}_A] &= \mathbb{E}[\tilde{p}_A + Z|\tilde{p}_A] \\ &= \mathbb{E}[\tilde{p}_A|\tilde{p}_A] + \mathbb{E}[Z|\tilde{p}_A] \\ &= \tilde{p}_A\end{aligned}$$

which is sufficient to show that $\tilde{P}_A$ is calibrated. To see why, recall that p_A is calibrated and note that by the tower property of conditional expectation with $\sigma(\tilde{p}_A) \subseteq \sigma(p_A)$,

$$\mathbb{E}[p_A|\tilde{p}_A] = \mathbb{E}[\mathbb{E}(Y|p_A)|\tilde{p}_A] = \mathbb{E}[Y|\tilde{p}_A]$$

□

Justification for discretizing p^*

Our algorithm uses the discretization of p^* to construct a linear program that maps probability masses from p^* to $\hat{p}$. Note that even if the original p^* is not discrete, it can easily be discretized into N bins by taking

$$p' = \lfloor Np^* \rfloor / N.$$

Note that the discretized score will satisfy $|p' - p^*| \leq N^{-1}$ almost surely, so for large values of N, the discretization p' well-approximates p^*.

Generalizing algorithm when p^* is not calibrated

The algorithm can be easily adapted for cases when the most accurate available estimate of p^* is not calibrated within groups. Part 1 of the algorithm (Section 3.1) remains unchanged as well as the need to perform Part 2 (Section 3.2) separately for each group A. But for each run of Part 2, a vector q needs to be computed from the discretized p^* and three constraints updated as follows.

For each discretized score assignment p_i of p^*, define q_i as the mean outcome of group-A individuals assigned p_i, i.e. $q_i = \mathbb{E}[Y|p^* = p_i, A]$. Then denote the vector of these conditional means as $q = (q_1, q_2, \ldots, q_N)$. Once q is computed, the core of the Part 2 algorithm is unchanged. We are still mapping the distribution s from the discretized scores p^* to the distribution f of the new fair score $\hat{p}$. However, we now use q in the following constraints that replace 17 and 18:

$$\sum_{i=1}^{N} T_{ij} q_i s_i = p_j f_j \quad \forall j \in \{1, \ldots, N\}. \tag{A.20}$$

$$\sum_{j=1}^{N} \sum_{i=1}^{N} T_{ij} q_i s_i \left(\mathbb{1}\{p_j \geq \bar{p}\} - \alpha_2\right) = 0 \tag{A.21}$$

$$\sum_{j=1}^{N} \sum_{i=1}^{N} T_{ij} (1 - q_i) s_i \left(\mathbb{1}\{p_j \geq \bar{p}\} - \alpha_1\right) = 0 \tag{A.22}$$

Generalizing algorithm given an interval of possible k or $\bar{p}$

In settings where the exact $\bar{p}$ is unknown or not fixed, our algorithm can be adapted for an interval of possible cutoffs $(\bar{p}-\epsilon, \bar{p}+\epsilon)$. The generalized version produces scores $\hat{p}$ that are either below $\bar{p}-\epsilon$ or above $\bar{p}+\epsilon$, so that any cutoff in the interval would execute the same classifications.

In particular, we propose a couple modifications to generalize our algorithm to this setting. We wish for anyone receiving scores above $\bar{p}+\epsilon$ to be classified as $\hat{y}=1$ and anyone receiving scores below $\bar{p}-\epsilon$ as $\hat{y}=0$. Following the reasoning in Proposition 2, for such a score to be calibrated, the associated PPV should exceed $\bar{p}+\epsilon$ and the NPV exceed $1-(\bar{p}-\epsilon)$. Therefore, the feasible region previously defined in Theorem 2 by 11 is now defined by

$$\frac{\alpha_2}{\alpha_1} \geq \frac{\bar{p}+\epsilon}{\beta_L(1-(\bar{p}+\epsilon))} \tag{A.23}$$

$$\frac{(1-\alpha_1)}{(1-\alpha_2)} > \frac{\beta_H(1-(\bar{p}-\epsilon))}{\bar{p}-\epsilon} \tag{A.24}$$

We also add a constraint to the scoring algorithm, specifying that no post-processed scores be assigned values inside the interval of possible cutoffs:

$$T_{ik} = 0 \quad \forall k \text{ such that } p_k \in (\bar{p}-\epsilon, \bar{p}+\epsilon) \tag{A.25}$$

The rest of the procedure remains unchanged. The cost of the added flexibility is a tighter feasible region and higher MSE of the final score.

Generalizing algorithm when there are more than two groups

This algorithm can be adapted to achieve the fairness criteria for multiple groups across multiple identifiers. First identify each unique group in A.

The feasible set of error rates is then all points in the intersection of each group's set $S(A)$ satisfying the inequalities (11) where H is the highest-mean group and L is the lowest-mean group. Then, as in Section 3.1, find the decision-maker's favored set of error rates in that feasible region.

Finally, for each group, separately implement the linear program in Section 3.2.

Appendix for Section 4: Credit Lending

Raw data and cleaning

Our empirical application is based on public data collected and made available by the U.S. Census Bureau, specifically the 2014 Survey of Income and Program Participation [1]. We converted the datasets from Waves 2 and 4 to CSV format and then organized them to serve our prediction task: use features in Wave 2 to predict reported repayment ability in Wave 4.

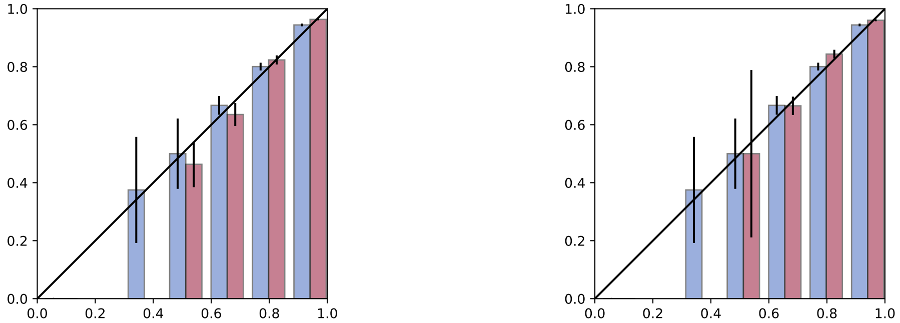
We matched every adult from the Wave 2 survey who responded to the Wave 4 survey and dropped the non-responders. We used education reported in Wave 2 to distinguish two groups L and H (representing 44% and 56% of the population respectively), L who attained at most a high school education and H who attained more. We randomly allocated 30% of all observations to a test set (about 8,000 adults) and the remaining 70% to a training set (about 18,000 adults).

Our outcome was the respondents' ability to pay mortgage, rent, and utilities in every month tracked in 2016 according to Wave 4. Any adult who failed to pay mortgage, rent, and/or utilities in any month was assigned label $Y=0$, and otherwise assigned $Y=1$. Base rates differed across groups; 11% of the less-educated group missed a payment compared to only 7% of the higher-educated group.

Finally, we constructed two sets of features. The first was based on rich data, comprising virtually all available variables from the Wave 2 survey but dropping those with no variation in the training set (leaving over 2,000 in total). The second was based on limited data, where we hand-selected "non-sensitive" variables involving assets, debts, income, and employment (over 800 in total).

[1] https://www.census.gov/programs-surveys/sipp/data/datasets.html

We identified which features were categorical and performed one-hot encoding. Then we standardized all features by centering them at 0 and dividing by their feature-specific standard deviations from the training set.



(a)

(b)

Figure A.4: Credit lending calibration plots. Panel A.4a depicts discretized pre-processed scores on the horizontal axis, with the portion in each bin paying their bills plotted on the vertical axis (including standard errors). Panel A.4b depicts the calibration of post-processed scores. Our procedure is seen to preserve calibration.

Deriving our empirical results

In deriving our empirical results, we employed the following protocol. As an initial step we estimated the original scores p^* with LASSO where the penalty parameters were tuned using 10-fold cross validation in the training dataset. We then tuned and evaluated the post-processing procedure in the test dataset. This consisted of the following steps.

1. We compute a discrete approximation to the score distribution of p^* for each group using the `numpy.histogram` Python method. This involves setting the user-defined hyperparameter N for the number of bins. We produced all results with the specification $N = 50$. We also tried $N = 10, 15, 25, 100, 250$, which did not appear to change results significantly. For $N = 500, 1000$, results also did not change significantly but the running time was significantly longer.
2. Next we calibrate the discrete approximation to the data by replacing the score assigned to each bin with the average outcome.
3. We use the calibrated and discretized scores to compute group-specific ROC curves using the `scikitlearn.metrics.roc_curve` function, and then compute the calibration compatibility constraints, assuming $k = 10 \implies \bar{p} = \frac{10}{11}$. These determine the feasible region $R(H) \times R(L)$.
4. We define the loss (14) as a function of error rates using $k = 10$, picking Λ to equate the true positive rates across groups. We minimize that loss in the feasible region using the `cvxpy` convex optimization library. We directly report the losses corresponding to the optima found by our procedure. Our comparisons correspond to removing
   (a) the calibration compatibility constraints,
   (b) the calibration compatibility constraints and the equal opportunity constraint
   (c) both constraints, as well as omitting sensitive features from estimation of p^*.
5. Then we use our post-processing method to back out the most informative score $\hat{p}$ that produces the optimal error rates. In particular, we compute the transformation kernel T using the `cvxpy` convex optimization library. Next we output post-processed scores by randomly mapping individuals' original scores given by p^* to new scores $\hat{p}$ with probabilities specified by the kernel T.

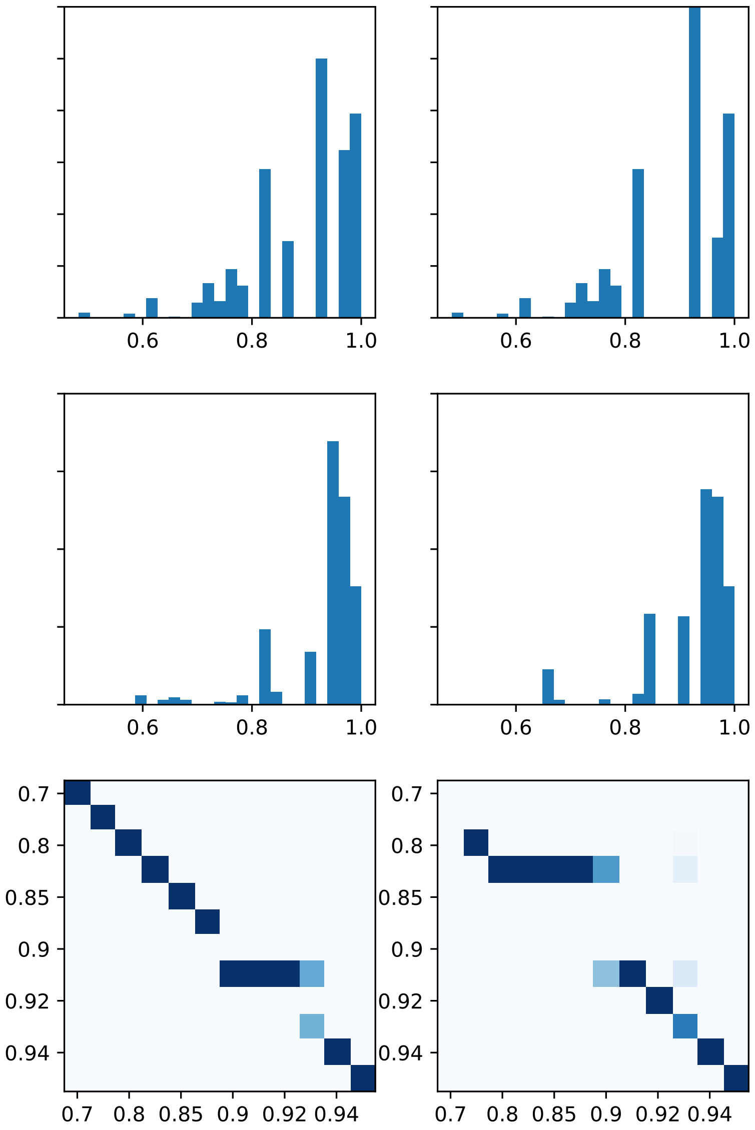


Figure A.5: Credit lending score comparisons. The top-most plots depict the distribution of scores in the less-educated group (inputted p^* on the left and outputted $\hat{p}$ on the right). The middle plots depict the distribution of scores in the high-educated group (inputted p^* on the left and outputted $\hat{p}$ on the right). The bottom plots depict how the post-processing procedure assigns probability masses from the inputted score (horizontal axis) to the outputted score (vertical axis), with the less-educated group's transformation depicted to the left and the high-educated group's transformation depicted to the right.

Finally, we evaluate our performance by inspecting the calibration of the output score and computing MSEs, error rates, and the decision-maker's loss. All losses we report are computed as a function of error rates, according to

$$\begin{aligned}\mathbb{E}\ell(\hat{y}, Y) &= \mathbb{E}[k\mathbb{1}\{\hat{y} > Y\} + \mathbb{1}\{\hat{y} < Y\}] \\ &= k\mathbb{P}(Y = 0)\mathbb{P}(\hat{y} = 1|Y = 0) + \mathbb{P}(Y = 1)\mathbb{P}(\hat{y} = 0|Y = 1).\end{aligned}$$

We simply replace the conditional probabilities by empirical averages from the test dataset. For the MSE of a risk score $\hat{p}$, we report $\mathbb{E}_n[(Y - \hat{p})^2]$ as is standard. Although not reported in the table, the standard deviation of the MSE of our (randomized) post-processed scores from 100 repetitions is 0.0001.

To assess the extent to which our post-processing preserves calibration, in Figure A.4 we plotted score bins on the horizontal axis and the average outcomes within each bin along the vertical axis. Error bars depict the standard error of the mean estimate within each bin.

We can also study how the post-processing transforms the most accurate estimates of p^* to the outputted scores $\hat{p}$ that satisfy the fairness criteria, Figure A.5 depicts in detail how the post-processing procedure shifts the original distribution of scores.

Appendix for Section 4: Criminal Justice

Raw data and cleaning

The second example in our paper shows that our procedure can modify existing risk assessments to output calibrated scores and corresponding binary summaries satisfying equal error rates. We used the Broward County dataset of COMPAS risk scores made available by *ProPublica*. [2]

Motivated by *ProPublica*'s analysis, we chose as our outcome the variable "two_year_recid" and supposed that COMPAS scores from 1-4 are classified as low risk while those from 5-10 are classified as high risk. We also considered only defendants labelled as white and black (40% and 60% respectively from a total sample of 6,150). Their recidivism rates vary. A percentage 51% of black defendants recidivated within two years, compared to 39% of the white defendants.

We define a positive label $Y = 1$ as *not* recidivating within two years, and otherwise assign label $Y = 0$. Defined as such, the white defendants in the dataset have a higher base rate than the black defendants.

Deriving our empirical results

Our results followed these steps.

1. We divide all defendants' given decile scores by 10 so they lie between 0 and 1.
2. Next we calibrate the group-specific scores by replacing each with the average outcome of individuals assigned that score.
3. We use the calibrated discrete scores to compute group-specific ROC curves using the `scikitlearn.metrics.roc_curve` function.
4. Then we back out the effective k and $\bar{p}$ that serve as inputs to the calibration compatibility constraint and the loss function. In particular, we wish to maintain the same effective risk cutoff in our post-processing as used in the *ProPublica* analysis on original COMPAS scores. Therefore for our purposes we define the cutoff $\bar{p}$ to be the minimum score (after calibration) that was classified in the *ProPublica* analysis as "high." Along with the group base rates, this determines the calibration compatibility constraints that combined with our ROC curves give the feasible region. Given the corresponding k, we define the loss function and find the optimal target rates in that region.
5. We use our risk score optimization method to back out the most informative scores that produce the target error rates. In particular, we compute the transformation kernel T using the `cvxpy` convex optimization library. Then we output post-processed scores by randomly

[2] The file "compas-scores-two-years.csv" is available at https://github.com/propublica/compas-analysis

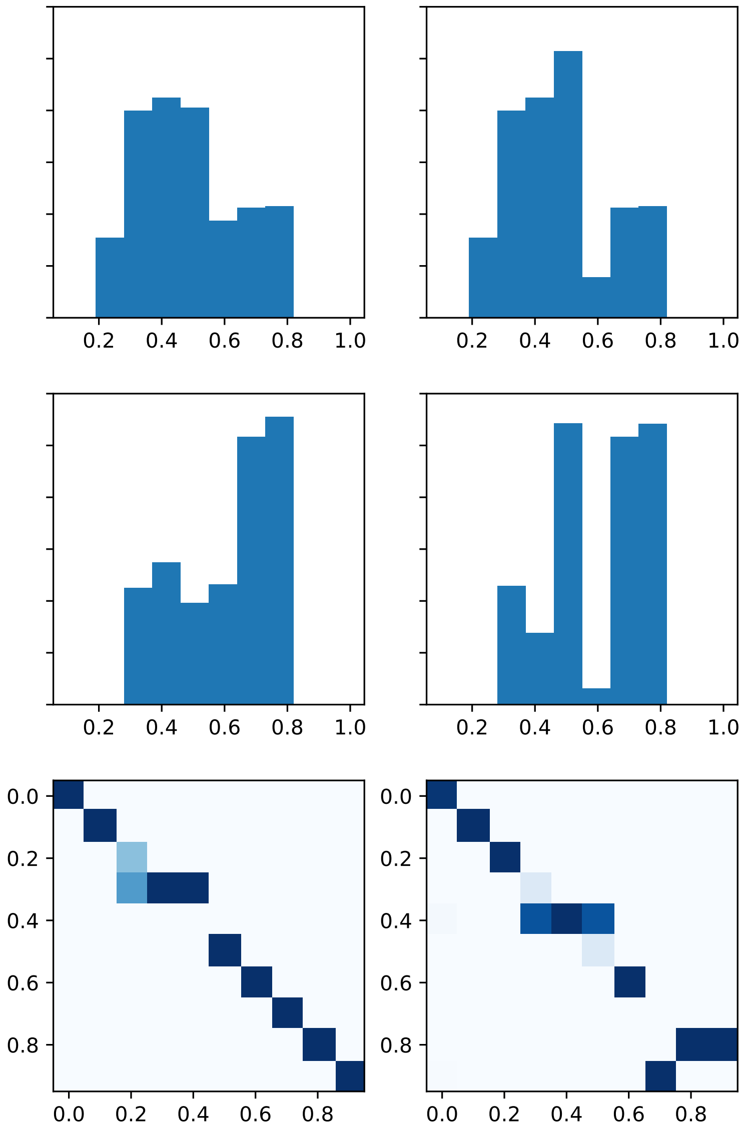


Figure A.6: Criminal justice score comparisons. Recall that we defined the scores to signify probabilities of *not* recidivating. The top-most plots depict the distribution of scores among black defendants in the dataset (inputted p^* on the left and outputted $\hat{p}$ on the right). The middle plots depict the distribution of scores among white defendants in the dataset (inputted p^* on the left and outputted $\hat{p}$ on the right). The bottom plots depict how the post-processing procedure assigns probability masses from the inputted score (horizontal axis) to the outputted score (vertical axis), with the black defendants' transformation kernel depicted to the left and the white defendants' transformation kernel depicted to the right.

mapping individuals' original scores given by p^* to new scores $\hat{p}$ with probabilities specified by the kernel T.

Finally, we plot the error rates that our post-processing achieves and compare them to the disparate rates found by *ProPublica*. We also produce a calibration plot showing that our procedure preserves predictive parity of the scores. To supplement the plots from the paper, Figure A.6 depicts how the post-processing procedure shifts the original distribution of scores to achieve the fairness criteria.