arXiv:2002.07676v1 [cs.LG] 18 Feb 2020

A Resolution in Algorithmic Fairness: Calibrated Scores for Fair Classifications

Claire Lazar [1] *
Suhas Vijaykumar [1]

Abstract

Calibration and equal error rates are fundamental conditions for algorithmic fairness that have been shown to conflict with each other, suggesting that they cannot be satisfied simultaneously. This paper shows that the two are in fact compatible and presents a method for reconciling them. 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 empirically to credit lending, our algorithm 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. These risk assessment tools hold the promise of driving immense social advancements, but they have also been shown to lead to unfair outcomes (Barocas & Selbst, 2016; Mitchell et al., 2019). If we can narrow these gaps of inequality in our prediction models, we will be better equipped to narrow them in the world.

However, producing fair risk assessments has proven to be challenging not only in practice but also in theory. Researchers have shown that even if a dataset is free from mismeasurement or discrimination, an algorithm trained on it will face significant fairness tradeoffs as long as groups represented in the data have different average outcomes (Kleinberg et al., 2016; Chouldechova, 2017; Berk et al., 2018; Corbett-Davies et al., 2017; Kleinberg & Mullainathan, 2019). 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 method to reconcile two important notions of fairness thought to be in conflict: calibration and equal error rates, which Hardt et al. (2016) call equalized odds. In influential work, these two criteria were proven to be mutually incompatible when both are applied to a *risk score* (Kleinberg et al., 2016; Pleiss et al., 2017) and when both are applied to a *classifier* (Chouldechova, 2017), suggesting that they may be incompatible altogether (Angwin, 2016a; Corbett-Davies et al., 2016).

We relax the mathematical tension between these two fairness criteria by enforcing *calibration on the score* and *equal error rates on the classifier*. In particular, we consider the classic setting of providing a risk score to a decision-maker, such as a lender, who then uses it to produce loss-minimizing binary classifications, such as loan approvals and denials. Our goal is to deliver calibrated scores that induce the decision-maker to make classifications satisfying equal error rates.

This formulation is not only theoretically convenient but also well-motivated from a practical point of view. First, our calibration condition on the score prompts the decision-maker to use a group-blind cutoff when making classifications. Second, our equal error rates condition ensures that an individual's probability of being correctly classified is independent of their group affiliation.

A key observation in both our theoretical and empirical analysis is that in our framework, *data richness complements fairness*. On the theoretical side, we show that the feasible space of fair solutions grows with the informativeness of the data; in our empirical application to credit lending, we demonstrate that using more features leads to higher accuracy and lower error rate disparity. Our results therefore contribute to the growing evidence that omission of sensitive features can be counter-productive to fairness (Dwork et al., 2012; 2018; Corbett-Davies & Goel, 2018; Kleinberg et al., 2018; Kleinberg & Mullainathan, 2019).

Our results proceed as follows. In Section 2, we prove that it is possible to construct calibrated scores that lead to equal

[1]MIT Institute for Data, Systems, and Society & Department of Economics, Cambridge, MA, United States.

* Correspondence to: Claire Lazar <clazar@mit.edu>.

error rate classifications and we precisely characterize when such a score exists. In Section 3, we propose an algorithm that produces the most informative score satisfying the fairness criteria and minimizing the decision-maker's errors. When applied in a setting that has no perfectly fair solution, our method instead yields calibrated scores that minimize error disparities across groups. Finally, in Section 4, we apply our proposed method to the problem of screening loan applicants. We compare our method to an alternative fairness rule that omits from the training data sensitive characteristics, such as demographic information, and is commonly used in practice. Our algorithm simultaneously achieves lower loss for the decision-maker and lower error rate disparity.

2. The Possibility of Fairness

2.1. Formalizing our Fairness Conditions

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 $\hat{p} = \hat{p}(X, A) \in [0, 1]$ 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 resulting classifications are based only on $\hat{p}$ and not on A, we constrain $\hat{p}$ to satisfy *calibration within groups*,

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

Since Y is binary, this implies that Y and A are conditionally independent given $\hat{p}$,

$$(Y \perp\!\!\!\perp A) \mid \hat{p}. \tag{4}$$

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

$$\mathbb{E}[\ell(Y, \hat{y})|\hat{p}, A] = \hat{p}(1 - \hat{y}) + k\hat{y}(1 - \hat{p}). \tag{5}$$

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

$$\hat{y} = \mathbb{1}\{\hat{p} \geq \bar{p}\}, \tag{6}$$

where $\bar{p} = {}^{k}/_{k+1}$ is fixed by the decision-maker.

Our second fairness condition constrains $\hat{y}$ to satisfy *equal error rates*, ensuring that the probability of correct classification is the same across groups:

$$(\hat{y} \perp\!\!\!\perp A) \mid Y. \tag{7}$$

Following the decision rule (6), we may write this as

$$(\mathbb{1}\{\hat{p} \geq \bar{p}\} \perp\!\!\!\perp A) \mid Y. \tag{8}$$

Our calibration and equal error rate conditions are summarized by (3) and (8), respectively. We refer to scores satisfying these properties as *fair*.

2.2. An Example from Credit Lending

We present an example to make our formalism concrete. Suppose we are tasked with estimating probabilities of credit repayment $\hat{p}$ to inform a lender's approvals $\hat{y}$ of loan applications. Our scores must fairly treat members of two groups, high-education applicants H and low-education applicants L. In particular, we wish for the lender's classifications to satisfy equal error rates: the probability that a good applicant will be accepted $\mathbb{P}(\hat{y} = 1|Y = 1, A)$ and a bad applicant denied $\mathbb{P}(\hat{y} = 0|Y = 0, A)$ should not depend on education level A.

Our objective is to determine whether there exist calibrated scores $\hat{p}$ that would execute equal error rate classifications $\hat{y}$ at the lender's given profit-maximizing cutoff $\bar{p}$, and if so, to find the most informative such score.

2.3. A General Impossibility Result and Our Departure

We first introduce a general impossibility result and show where our assumptions diverge to make fairness possible. 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) $(Y \perp\!\!\!\perp A) \mid Z$,

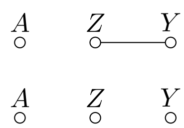
(ii) $(Z \perp\!\!\!\perp A) \mid Y$,

(iii) $\mathrm{Var}(A, Y|Z)$ is almost surely positive definite.

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

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 slightly stronger than requiring no perfect prediction and can be relaxed. Thus the theorem shows that when there is predictive uncertainty and Y depends on A (i.e. when the base rates are unequal), then it is impossible for a single Z to satisfy both calibration and equal error rates.

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, the conditional independence statements 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. $\square$

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 $\hat{y} = \mathbb{1}\{\hat{p} \geq \bar{p}\}$.

2.4. Necessary and Sufficient Conditions

In this section we characterize fairness feasibility, which we define as the existence of 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 input features X.

The graphical framework in this section builds on methods developed in Hardt et al. (2016). 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. A fairness-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.

2.4.1. 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. (2016) 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 $\mathbb{R}^2$,

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

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

$$S(A) = \{\alpha(\hat{y}, A) \,|\, \hat{y} = h(X, A), h \in \mathcal{H}\} \subseteq \mathbb{R}^2, \qquad (9)$$

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 describe $S(A)$ using Receiver Operator Characteristic (ROC) curves following Hardt et al. (2016). 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.

Lemma 1. $S(A)$ is convex.

This lemma supports the following proposition.

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

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

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

2.4.2. 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

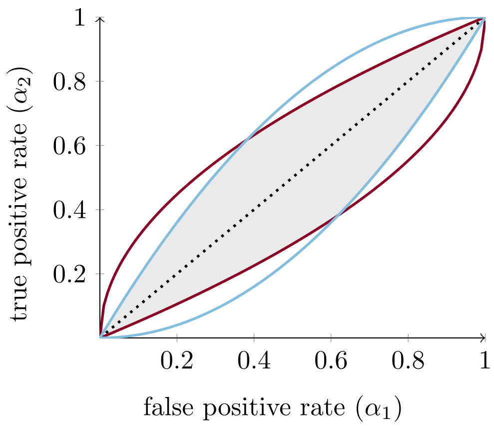


Figure 1. The space of equal error rate classifications. Depicted are two pairs of ROC curves that form the boundaries of $S(L)$ and $S(H)$. The shaded intersection $S(L) \cap S(H)$ marks the error rates associated with feasible equal error rate classifiers.

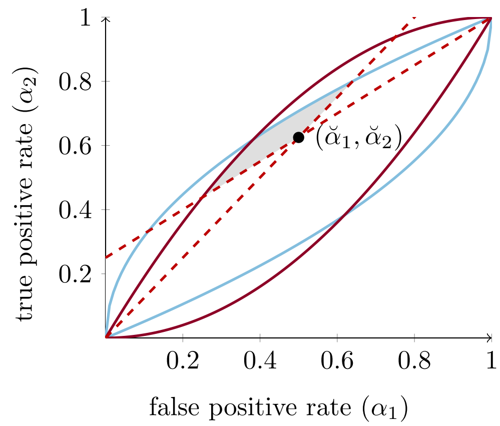


Figure 2. The fairness feasibility region (shaded). Restrictions (14) and (15) correspond to half-spaces above the red dashed lines.

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, $\forall A \in \{L, H\}$,

$$\mathbb{P}(Y = 1 | \hat{y} = 1, A) \geq \bar{p} \tag{10}$$

$$\mathbb{P}(Y = 0 | \hat{y} = 0, A) > 1 - \bar{p}. \tag{11}$$

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

$$\begin{aligned} \mathbb{P}(Y = 1 | \hat{y} = 1, A) &= \mathbb{E}[Y | \hat{p} \geq \bar{p}, A] \\ &= \mathbb{E}[\hat{p} | \hat{p} \geq \bar{p}, A] \geq \bar{p}. \end{aligned} \tag{12}$$

Similarly, it must also satisfy

$$\mathbb{P}(Y = 1 | \hat{y} = 0, A) = \mathbb{E}[\hat{p} | \hat{p} < \bar{p}, A] < \bar{p}. \tag{13}$$

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 (12) and (13), 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 $\hat{y} = \mathbb{1}\{\hat{p} \geq \bar{p}\}$. □

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

2.4.3. The Fairness Feasibility Region

Proposition 2 demonstrates that the following are equivalent:

(i) Fairness is feasible; 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, (10), and (11).

In practice, we propose checking (ii) to identify whether (i) holds. To do so, we will use Bayes' rule to write (10) and (11) 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. (2016). The following theorem and the accompanying Figure 2 indicates that each restriction (10) and (11) 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

$$\frac{\alpha_2}{\alpha_1} \geq \frac{\bar{p}}{\beta_L(1 - \bar{p})} \tag{14}$$

$$\frac{(1 - \alpha_1)}{(1 - \alpha_2)} > \frac{\beta_H(1 - \bar{p})}{\bar{p}} \tag{15}$$

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 (14) and (15) 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 $\bar{p}$,

$$\breve{\alpha}_1 = \frac{\beta_L}{(\beta_H - \beta_L)} \left(\frac{\beta_H - (1+\beta_H)\,\bar{p}}{\bar{p}} \right) \tag{16}$$

$$\breve{\alpha}_2 = \frac{1}{(\beta_H - \beta_L)} \left(\frac{\beta_H\,(1-\bar{p}) - \bar{p}}{1-\bar{p}} \right). \tag{17}$$

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 (14) and (15) 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 Fair Loss-Minimizing Algorithm

After checking for fairness feasibility, 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 the decision-maker's 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 cutoff $\bar{p}$ (Section 3.2).

3.1. Optimal Error Rates

We first describe how we identify the error rates that minimize loss subject to fairness feasibility. Let R denote the set of points (α_1, α_2) in the fairness-feasible region, i.e. the points in $S(H) \cap S(L)$ that satisfy inequalities (14) and (15). We separately consider the case where R is nonempty, and then formulate a flexible convex program that minimizes error disparity even if it is empty.

3.1.1. Assuming R is nonempty

R is necessarily convex, as it is the intersection of four convex regions: $S(H)$, $S(L)$, and the half-spaces satisfying (14) and (15). 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

$$\ell(\alpha_1, \alpha_2) \equiv k\alpha_1(1 - \mathbb{E}[Y]) + (1 - \alpha_2)\mathbb{E}[Y]. \tag{18}$$

over $(\alpha_1, \alpha_2) \in R$. The optimal (α_1, α_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 relative preference k for false positive and false negative classifications. We summarize this procedure below as 1A.

Algorithm Part 1A Optimal equal error rates

Input: Scores $\{p_i^*\}$, labels $\{Y_i\}$, group identities $\{A_i\}$, cutoff $\bar{p}$.
Step 1: Compute group-wise ROC curves and constraints (14) and (15). These determine the region R.
Step 2: Determine the point $(\breve{\alpha}_1, \breve{\alpha}_2)$ according to (16) and (17).
if $\breve{\alpha}_1 \in (0, 1)$ and $(\breve{\alpha}_1, \breve{\alpha}_2)$ is above any of top two ROC curves **then**
  Output: No feasible solution.
end if
Step 3: Minimize (18) over $(\alpha_1, \alpha_2) \in R$.
Output: Optimal (α_1^*, α_2^*) from Step 3.

3.1.2. General formulation

We next present a general formulation that also accommodates the case where R is empty, in which case we penalize error rate discrepancies between the two groups among classifiers compatible with calibration.

Furthermore, the general formulation provides the flexibility of varying the relative importance λ of minimizing groups' error rate differences and the decision-maker's loss. In settings where overall accuracy loss is highly costly, such as in medical assessments (Chen et al., 2018), a small λ can be chosen and our second phase of the algorithm to find a calibrated corresponding $\hat{p}$ can still be applied.

First define $R(A)$ as the set of true and false positive rates in $S(A)$ achievable from thresholding a calibrated score at $\bar{p}$, i.e. satisfying (14) and (15). Specifically,

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

Letting γ denote the fraction of individuals in group L, the revised convex program finds $z_L^* = (\alpha_{1L}^*, \alpha_{2L}^*) \in R(L)$

and $z_H^* = (\alpha_{1H}^*, \alpha_{2H}^*) \in R(H)$ that minimize the expression

$$\gamma\ell(z_L) + (1-\gamma)\ell(z_H) + \lambda \left\| z_L - z_H \right\|_2 \qquad (19)$$

Note that when R is nonempty and λ is arbitrarily large, then a solution to (19) produces equal group-specific error rates, $z_L^* = z_H^*$.

Algorithm Part 1B Penalized error discrepancy

Input: Scores $\{p_i^*\}$, labels $\{Y_i\}$, group identities $\{A_i\}$, penalty λ, cutoff $\bar{p}$.
Step 1: Compute group-wise ROC curves and constraints (14) and (15). These determine the regions $R(H)$ and $R(L)$.
Step 2: Minimize (19) over $(z_L, z_H) \in R(L) \times R(H)$.
Output: Optimal (z_L^*, z_H^*) from Step 2.

3.2. Optimal Risk Scores

Once the optimal feasible error rates z_L^* and z_H^* are chosen, the decision-maker's expected loss becomes determined. However, multiple choices of calibrated scores may be compatible with those target rates at the cutoff $\bar{p}$, and likely some scores are more informative than others. This section describes a method to recover the most accurate score $\hat{p}$ that implements the error rates z_L^* and z_H^* by solving a constrained optimal transport problem (Peyré et al., 2019).

We base the method on the observation that the best fair $\hat{p}$ is recoverable through post-processing the Bayes optimal score $p^* = \mathbb{E}[Y|X, A]$. In their development of an equal error rate classifier, Hardt et al. (2016) likewise justify post-processing, and we adapt their argument to our fairness setting in the supplement. For clarity of presentation we will assume the likely fact that p^* already satisfies calibration within groups (Liu et al., 2019). This is relaxed in the supplement.

Finally, in the supplement we prove that any arbitrary post-processing of a calibrated p^* that yields a calibrated $\hat{p}$ is a mean-preserving contraction. Viewed this way, our method can be thought of as searching for the smallest mean-preserving contraction of p^* that implements the decision-maker's preferred solution.

3.2.1. Proposed Scoring Algorithm

We define a separate linear program per group A and seek 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 $\hat{p}$. To make the problem tractable, we discretize p^* 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^*. We assume for simplicity that p^* itself is discrete. It takes N ordered values $p = (p_1, p_2, \ldots, p_N)$, each with probability mass given by $s = (s_1, s_2, \ldots, s_N)$ 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, $f = (f_1, f_2, \ldots, f_N)$.

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

$$T's = f. \qquad (20)$$

In order to produce probability distributions, elements of T must take values between 0 and 1, and each of its rows should sum to 1,

$$0 \leq T_{ij} \leq 1 \text{ and } \sum_{k=1}^{N} T_{ik} = 1 \quad \forall\, i, j \in \{1, \ldots N\}. \quad (21)$$

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. Under our assumption that p^* is calibrated, this reduces to

$$\sum_{i=1}^{N} T_{ij} p_i s_i = p_j f_j \quad \forall\, j \in \{1, \ldots, N\}. \qquad (22)$$

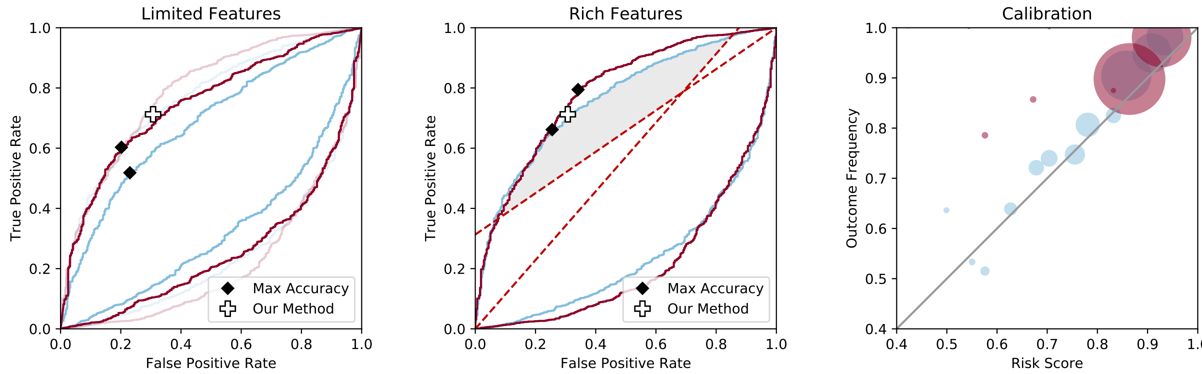
In addition, the targeted false- and true-positive rates (α_1, α_2) derived in Section 3.1 give the following constraints:

$$\sum_{j=1}^{N} \sum_{i=1}^{N} T_{ij} p_i s_i \left(\mathbb{1}\{p_j \geq \bar{p}\} - \alpha_2 \right) = 0, \qquad (23)$$

$$\sum_{j=1}^{N} \sum_{i=1}^{N} T_{ij} (1 - p_i) s_i \left(\mathbb{1}\{p_j \geq \bar{p}\} - \alpha_1 \right) = 0. \qquad (24)$$

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

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



(a) (b) (c)

Figure 3. Illustration of our algorithm and comparison with data omission strategy. Maroon indicates the H group; blue indicates the L group. Panel (a) illustrates the empirical ROC curves derived from the limited feature set with the decision-makers optimal solutions. Our solution and corresponding ROC curves are overlaid for comparison. Panel (b) illustrates the empirical ROC curves derived from the rich feature set, with the most accurate scores yielding indicated optimal solutions. Our post-processing method bridges the two groups error rates. Panel (c) shows our scores are calibrated within groups as most lie on the main diagonal.

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 score that minimizes mean-squared error will also minimize

$$\mathbb{E}[(\hat{p}-p^*)^2] = \sum_{i=1}^{N}\sum_{j=1}^{N} T_{ij}(p_i - p_j)^2 s_i. \qquad (25)$$

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

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

Thus, by minimizing the objective (25) we can effectively control the additional error incurred by post-processing. Combining this objective with our constraints, we arrive at an algorithm that solves a constrained, one-dimensional optimal transportation problem with quadratic cost.

Algorithm Part 2 Post-processed Scores

Input: Group-specific scores p^*, number of bins N, target error rates $z^* = (\alpha_1^*, \alpha_2^*)$, cutoff $\bar{p}$.
Step 1: Produce discrete score approximation $(p_1, p_2, \ldots, p_N)$ and $(s_1, s_2, \ldots, s_N)$.
Step 2: Choose T to minimize (25) subject to the constraints (20), (21), (22), (23) and (24).
Step 3: Produce $\hat{p}$ by mapping each individual input score q to p_j with independent probability $T_{\lfloor Nq \rfloor, j}$.
Output: Group-specific scores $\hat{p}$.

In the following section, we take our procedure to the data.

4. Empirical Results

We now return to our previous example of designing a risk score to inform a credit lenders approvals of loan applicants, which we conceptualized in Section 2.2. For our example we use the U.S. Census Bureau's Survey of Income and Program Participation (SIPP), a nationally-representative survey of the civilian population with thousands of features. We select as our outcome the reported ability to pay rent, mortgage, and utilities in 2016, and predict the outcome using the survey responses from two years prior. We distinguish two groups by education level. Of the overall population, 56% have studied beyond a high school education and are labelled H, with the rest of the population labelled L. The base rates of these two groups are different: 11% of the less-educated group L missed a payment in 2016, compared to 7% of the H group. We assume the lender views missed payments as costly and has optimal cutoff $\bar{p} \approx .9$.

We construct two collections of features: one is *limited*, containing about 800 detailed financial variables relating to work history, assets, and debts; the other is *rich*, using all of these plus over 1000 additional variables spanning aspects of food security, family structure, and participation in social programs, among others. Intuitively, it may seem that the limited feature set would produce outcomes that are costlier for the lender but fairer to applicants, as it is restricted to financial variables and excludes detailed information about individuals' education and non-financial characteristics. However, we find that the limited feature set produces a solution worse for both the lender and the applicants. Our results are summarized numerically in Table 1 and graphically in Figure 3 using a test set that makes up 30% of our observations.

Table 1. Predictions using both limited and rich feature sets. Rows represent types of algorithmic outputs supplied to the lender, and columns are based on the lender's simulated classifications given those outputs. Row [1] is based on our algorithm using all features, enforcing calibration and equal error rates. Row [2] implements the algorithm of Hardt et al. (2016) to derive equal error rate classifications, but not a score. Since it retrieves the same outcomes as [1] we see there is no additional loss from enforcing calibration. Row [3] uses all features and does not post-process, leading to disparate error rates. Row [4] uses a restricted feature set and displays higher loss for the lender, lower true positive rates for both groups, and substantial error disparities across groups.

	Algorithm	Lender Loss	TPR (H/L)	FPR (H/L)	Score MSE	Blind Cutoff
	All Features					
[1]	*Eq. Errors + Calibration*	.542	(.713/.713)	(.307/.307)	.073	✓
[2]	Eq. Errors Only	.542	(.713/.713)	(.307/.307)	N/A	
[3]	Accuracy Maximizing	.517	(.795/.661)	(.341/.255)	.071	
	Limited Features					
[4]	Accuracy Maximizing	.591	(.603/.518)	(.202/.230)	.077	

For each feature set we use cross-validated LASSO to estimate a separate set of risk scores p^*. We then compute group-specific empirical ROC curves, which are illustrated in Figures 3a and 3b. Consider first models based on the limited feature set (Figure 3a). As expected, they are associated with lower ROC curves for both groups compared to those derived from the rich feature set (Figure 3b). What is less intuitive is that despite the removal of sensitive training data, less-educated individuals in Figure 3a have substantially *lower* ROC curves than their more-educated counterparts. This indicates that algorithms trained on the limited data would struggle to correctly predict outcomes for the L group in particular. We observe three results from the accuracy-maximizing model that omits sensitive features: lenders face higher loss, there is substantial error rate disparity across education groups, and creditworthy applicants from both education groups are less likely to be granted a loan. Furthermore, there is significant incentive for the lender to choose group-dependent cutoffs.

These results indicate a counter-productive side effect of omitting sensitive features from algorithms. On one hand, omitting sensitive features is intended to affect prediction rankings across groups and give a helpful boost to applicants from protected classes. However, the approach typically also weakens accuracy *within* groups, harming both the decision-maker as well as creditworthy applicants that can no longer be distinguished from defaulters. This phenomenon has been discussed by Kleinberg et al. (2018), who challenge the idea that omission of data is the path to fairness.

In comparison, we present our algorithm in Figure 3b. The shaded region indicates the feasible equal error rates achievable by a calibrated score at $\bar{p}$, the decision maker's favored cutoff. We indicate with a cross the decision-maker's optimal choice in this region. Then, using our proposed algorithm, we derive a score that achieves those targeted error rates. Figure 3c plots predicted probabilities of repayment on the horizontal axis and the actual incidence of repayment on the vertical, showing that both groups' scores are calibrated. From the perspective of both lenders and applicants, our method outperforms the strategy of omitting data to achieve fair outcomes.

5. Conclusion

In settings from hospitals to courts, decision-makers stand to benefit from algorithmic predictions. This paper has presented a path toward fair prediction in the widespread setting where a risk score is constructed to aid their classification tasks. We prove that although the fairness conditions of calibration and equal error rates are thought to be in conflict, in fact it can be possible to construct calibrated scores that lead to equal error rate classifications. We characterize exactly when a solution is possible and propose an algorithm that creates 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 an algorithm based on only non-sensitive financial data to predict credit-worthiness, our algorithm trained on an unrestricted set of features produces a solution that enhances both efficiency and equity.

Acknowledgements

Many thanks to Iván Werning, Anna Mikusheva, 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 and Adam Fisch 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

Angwin, Julia, J. L. 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 2016a.

Angwin, Julia, J. L. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, May 2016b.

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

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, pp. 0049124118782533, 2018.

Chen, I., Johansson, F. D., and Sontag, D. Why Is My Classifier Discriminatory? *arXiv:1805.12002 [cs, stat]*, December 2018.

Chouldechova, A. 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.

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

Corbett-Davies, S., Pierson, E., Feller, A., and Goel, S. 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.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. 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, pp. 797–806, Halifax, NS, Canada, 2017. Association for Computing Machinery. ISBN 978-1-4503-4887-4. doi: 10.1145/3097983.3098095.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pp. 214–226, Cambridge, Massachusetts, 2012. Association for Computing Machinery. ISBN 978-1-4503-1115-1. doi: 10.1145/2090236.2090255.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. Decoupled classifiers for group-fair and efficient machine learning. In Friedler, S. A. and Wilson, C. (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 119–133, New York, NY, USA, February 2018. PMLR.

Hardt, M., Price, E., Price, E., and Srebro, N. Equality of Opportunity in Supervised Learning. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 3315–3323. Curran Associates, Inc., 2016.

Kleinberg, J. and Mullainathan, S. Simplicity creates inequity: Implications for fairness, stereotypes, and interpretability. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19, pp. 807–808, Phoenix, AZ, USA, 2019. Association for Computing Machinery. ISBN 978-1-4503-6792-9. doi: 10.1145/3328526.3329621.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv:1609.05807 [cs, stat]*, November 2016.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Rambachan, A. Algorithmic Fairness. *AEA Papers and Proceedings*, 108:22–27, 2018. ISSN 2574-0768. doi: 10.1257/pandp.20181018.

Liu, L. T., Simchowitz, M., and Hardt, M. The implicit fairness criterion of unconstrained learning. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4051–4060, Long Beach, California, USA, June 2019. PMLR.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. *arXiv:1811.07867 [stat]*, July 2019.

Peyré, G., Cuturi, M., et al. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6): 355–607, 2019.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On Fairness and Calibration. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5680–5689. Curran Associates, Inc., 2017.

Villani, C. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2008.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pp. 1171–1180, Perth, Australia, 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

2.4.1

Proof of Lemma 1. Let ξ be an independent $\text{Ber}(\lambda)$ random variable. Then, by iterating expectations, one sees that

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

□

Proof of Proposition 1. 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 Lemma 1. For further details, 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 $S(A)$ (by Lemma 1) 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 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. □

2.4.3

Extension of Proposition 2. 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 (12) and (13), which are equivalent to (10) and (11) respectively. 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, (10), and (11). 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. Building on the 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 (14) and (15) is equivalent to the following: There exists a classifier $\hat{y}$ satisfying equal error rates, (10), and (11).

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 conditions (10) and (11) $\forall A$ is equivalent to its corresponding true and false positive rates $(\alpha_1(\hat{y}_e, A), \alpha_2(\hat{y}_e, A))$ satisfying (14) and (15) $\forall A$.

Recall that (10) 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 (14). We next similarly transform (11), 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 (15).

Note that special attention should be given to the corner solutions. At point $(0,0)$, (14) becomes irrelevant and so (15) is necessary and sufficient. Meanwhile at $(1,1)$, (15) becomes irrelevant so (14) is necessary and sufficient. □

Proof of Corollary 1. Let F and G denote the lines for which the inequalities (14) and (15) hold with equality. That is to say, $F, G \subset \mathbb{R}^2$ are given by

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

The lines intersect at $(\breve{\alpha}_1, \breve{\alpha}_2)$ given by (16) and (17). 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 $\breve{\alpha}_1 \le 0$, $\breve{\alpha}_1 \ge 1$, 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 (14) and (15). 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 (14) and (15) hold at $(\breve{\alpha}_1, \alpha_2)$. This is a feasible point.

Case II: $\breve{\alpha}_1 \le 0$. On the other hand, if $\breve{\alpha}_1 \le 0$, 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 (15) holds and the point is feasible.

Case III: $\breve{\alpha}_1 \ge 1$. If $\breve{\alpha}_1 \ge 1$, 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 (14) 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 $S(L)\cap S(H)$ and the positive slope of 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 the procedure to post-process 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, $\mathbb{E}[\ell(\mathbb{1}_{\hat{p}>\underline{p}}, Y)] \le \mathbb{E}[\ell(\mathbb{1}_{\hat{g}>\underline{p}}, Y)]$ 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 $\mathbb{1}\{p \ge \bar{p}\}$ 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 equal 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 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 (1) is satisfied with equality. $\square$

Mean-preserving contractions of calibrated scores

Next we describe our observation that calibrated scores derived from each other are related by mean-preserving spreads and contractions, with proof.

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 is a mean-preserving contraction of p_A with $p_A = \tilde{p}_A + Z$ and $\mathbb{E}[Z|\tilde{p}_A] = 0$ 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 $\sigma(\tilde{p}_A) \subseteq \sigma(p_A)$,

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

$\square$

Generalizing algorithm when p^* is not calibrated

The algorithm can be easily adapted for cases when the most accurate estimate of p^* is not calibrated within groups. Part 1 of the algorithm remains unchanged as well as the need to perform Part 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 three constraints that replace 22, 23, and 24:

$$\sum_{i=1}^{N} T_{ij} q_i s_i = p_j f_j \quad \forall j \in \{1, \ldots, N\}. \tag{A.26}$$

$$\sum_{j=1}^{N} \sum_{i=1}^{N} T_{ij} q_i s_i \left(\mathbb{1}\{p_j \geq \bar{p}\} - \alpha_2\right) = 0 \tag{A.27}$$

$$\sum_{j=1}^{N} \sum_{i=1}^{N} T_{ij} (1 - q_i) s_i \left(\mathbb{1}\{p_j \geq \bar{p}\} - \alpha_1\right) = 0 \tag{A.28}$$