From Soft Classifiers to Hard Decisions: How fair can we be?*

Ran Canetti†
Boston University and
Tel Aviv University
canetti@bu.edu

Aloni Cohen‡
MIT
aloni@mit.edu

Nishanth Dikkala§
MIT
nishanthd@csail.mit.edu

Govind Ramnarayan¶
MIT
govind@mit.edu

Sarah Scheffler‖
Boston University
sscheff@bu.edu

Adam Smith**
Boston University
ads22@bu.edu

ABSTRACT

A popular methodology for building binary decision-making classifiers in the presence of imperfect information is to first construct a calibrated non-binary "scoring" classifier, and then to post-process this score to obtain a binary decision. We study various fairness (or, error-balance) properties of this methodology, when the non-binary scores are calibrated over all protected groups, and with a variety of post-processing algorithms. Specifically, we show:

First, there does not exist a general way to post-process a calibrated classifier to equalize protected groups' positive or negative predictive value (PPV or NPV). For certain "nice" calibrated classifiers, either PPV or NPV can be equalized when the post-processor uses different thresholds across protected groups. Still, when the post-processing consists of a single global threshold across all groups, natural fairness properties, such as equalizing PPV in a nontrivial way, do not hold even for "nice" classifiers.

Second, when the post-processing stage is allowed to *defer* on some decisions (that is, to avoid making a decision by handing off some examples to a separate process), then for the non-deferred decisions, the resulting classifier can be made to equalize PPV, NPV, false positive rate (FPR) and false negative rate (FNR) across the protected groups. This suggests a way to partially evade the impossibility results of Chouldechova and Kleinberg et al., which preclude equalizing all of these measures simultaneously. We also present different deferring strategies and show how they affect the fairness properties of the overall system.

We evaluate our post-processing techniques using the COMPAS data set from 2016.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning*.

KEYWORDS

algorithmic fairness, classification, post-processing

ACM Reference Format:
Ran Canetti, Aloni Cohen, Nishanth Dikkala, Govind Ramnarayan, Sarah Scheffler, and Adam Smith. 2019. From Soft Classifiers to Hard Decisions: How fair can we be?. In *FAT* '19: Conference on Fairness, Accountability, and Transparency (FAT* '19), January 29–31, 2019, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3287560.3287561

1 INTRODUCTION

The concept of *fairness* is deeply ingrained in our psyche as a fundamental, essential ingredient of Human existence. Indeed the perception of fairness, broadly construed as accepting each others' equal right for well being, is arguably one of the most basic tenets of cooperative societies of individuals in general.

However, as fundamental as this concept may be, it is also elusive: different cultures have developed very different notions of fairness and equality among individuals, subject to religious, ethical, and social beliefs; in particular, the intricate interplay between fairness and justice is often left to subjective interpretation.

In the context of decision processes, fairness is further complicated by the fact that decisions are often made with *incomplete information* and *limited resources*. These two factors have become increasingly prominent as society grows and decision processes become more complex and algorithmic.

One way that researchers are responding to these growing concerns is by attempting to formulate precise notions for *fairness* of decisions processes, e.g. [5, 7, 12]. While these notions do not intend to capture the complexities of the ethical, socio-economic, or religious aspects of fairness, they do consider the fairness aspects of statistical decision-making processes with incomplete information. Essentially, these notions accept the fact that a decision process with incomplete information will inevitably make errors relative to the desired full-information notion (which is treated as a given), and provide guidelines on how to "balance the errors fairly" across

*The order of authors is alphabetical and does not convey any information on the authors' relative contributions. The authors are thankful for the helpful feedback of the anonymous reviewers. The full version of this paper can be found at [3].
†Member of CPIIS. Supported by NSF awards 1413920 & 1801564, ISF award 1523/14.
‡Supported by NSF award CNS-1413920, the 2018 Facebook Fellowship, and MIT's RSA Professorship and Fintech Initiative.
§Supported by NSF award CCF-1617730, CCF-1650733, and ONR N00014-12-1-0999.
¶Supported by NSF award CCF-1665252 and NSF award DMS-1737944.
‖Supported by the Clare Boothe Luce Graduate Research Fellowship and NSF award 1414119.
**Supported in part by NSF awards IIS-1447700 and AF-1763786 and a Sloan Foundation Research Award.

FAT '19, January 29–31, 2019, Atlanta, GA, USA*

ACM ISBN 978-1-4503-6125-5/19/01…$15.00
https://doi.org/10.1145/3287560.3287561

individuals or groups of individuals. These definitions have proven to be meaningful and eye opening; in particular, it has been demonstrated that some very natural measures of "fair distribution of error" are mutually inconsistent: No decision mechanism with incomplete information can satisfy all, except in trivial cases [4, 12].

Faced with this basic impossibility, we aim to better understand the process of decision making with incomplete information, and propose ways to relax the known measures so as to regain feasibility.

Specifically, we concentrate on the task of post-processing a calibrated soft classifier to obtain a binary decision, under group fairness constraints, for the case of several disjoint *protected groups*.

That is, we consider the following two-stage mechanism. The first stage consists of constructing a classifier $\hat{S}$ that outputs for each individual x a score $s \in [0, 1]$ that is related to the chance that x has property B. The only requirement we make of $\hat{S}$ is group-wise calibration: For each group g and for each $s \in [0, 1]$, the fraction of individuals in g that get score s and have the property, out of all individuals in g that get score s, is s. The second stage takes as input the output $s = \hat{S}(x)$ of the first stage and the group to which x belongs, and outputs a binary decision, interpreted as its guess at whether x has property B.

The first stage is aimed at gathering information and providing the best accuracy possible, with only minimal regard to fairness (i.e only group-wise calibration). The second stage extracts a decision from the information collected in the first stage, while making sure that the errors are distributed "fairly."

To further focus our study, we take the first stage as a given and concentrate on the second. That is, we consider the problem of *post-processing* the scores given by the calibrated soft classifier $\hat{S}$ into binary predictions. A representative example is a judge making a bail decision based on a score provided by a software package. Following [4, 9], we consider the following four performance measures for the resulting binary classifier: the *positive predictive value (PPV)*, namely the fraction of individuals that have the property among all individuals that the classifier predicted to have the property; The *false positive rate (FPR)*, namely the fraction of individuals that were predicted to have the property among all individuals that don't have the property; The *negative predictive value (NPV)* and *false negative rate (FNR)*, which are defined analogously. Ideally, we would like to equalize each one of the four measures across the groups, i.e. the measure will have the same value when restricted to samples from each group. Unfortunately, however, we know that this is impossible in general [4, 12]. This leads us to a broad question that motivates our work:

> Under what conditions can we post-process a calibrated soft classifier's outputs so that the resulting hard classifier equates a subset of {PPV, NPV, FNR, FPR} across a set of protected groups? How can we balance these conflicting goals?

Results: Post-Processing With Thesholds. In a first set of results we consider the properties obtained by post-processing via a "threshold" mechanism. Naively, a threshold post-processing mechanism would return 1 for individual x whenever the calibrated score $s(x)$ is above some fixed threshold, and return 0 otherwise. We somewhat extend this mechanism by allowing the post-processor "fine-tune" its decision by choosing the output probabilistically whenever the result of the soft classifier is exactly the threshold.

We first observe that the popular and natural pos-t-procesing method of using a single threshold across all groups has some inherent deficiency: No such mechanism can in general guarantee equality of either PPV or NPV across the protected groups.

We then show that, when using different thresholds for the different groups, one can equalize *either* PPV or NPV (but not both) across the two groups, assuming the profile of $\hat{S}$ has some non-degeneracy property.

The combination of the impossibility of single threshold and the possibility of per-group threshold also stands in contrast to the belief that a soft classifier that is calibrated across both groups allows "ignoring" group-membership information in any post-processing decision [14]. Indeed, the conversion to a binary decision "loses information" in different ways for the two groups, and so group membership becomes relevant again after post-processing.

Results: Adding deferrals. For the second set of results we consider post-processing strategies that do not always output a decision. Rather, with some probability the output is $\perp$, or "I dont know", which means that the decision is deferred to another (hopefully higher quality, even if more expensive) process. Let us first present our technical results and then discuss potential interpretations and context.

The first strategy is a natural extension of the per-group threshold: we use *two* thresholds per group, returning 1 above the right threshold, 0 below the left threshold, and $\perp$ between the thresholds. We show that there always exists a way to choose the thresholds such that, conditioned on the decision not being $\perp$, both the PPV and NPV are equal across groups.

Next we show a family of post-processing strategies where, conditioned on the decision not being $\perp$, *all four quantities* (PPV, NPV, FPR, FNR) are equal across groups.

All strategies in this family have the following structure: Given an individual x, the strategy first makes a randomized decision whether to defer on x, where the probability depends on $\hat{S}(x)$ and the group membership of x. If not deferred, then the decision is made via another post-processing technique.

One method for determining the probabilities of deferral is to make sure that the profiles of scores returned by the calibrated soft classifier, *conditioned on not deferring*, is equal for the two groups (That is, let $p_{s,g}$ denote the probability, restricted to group g, that an element gets score s conditioned on not deferring. Then for any s, we choose deferral probabilities so that $p_{s,g_1} = p_{s,g_2}$.) The resulting classifier can then be post-processed in *any* group-blind way (say, via a single threshold mechanism as described above).

Of course, the fact that all four quantities are equalized conditioned on not deferring does not, in and of itself, provide any guarantees regarding the fairness properties of the overall decision process — which includes also the downstream decision mechanism. For one, it would be naive to simply assume that fairness "composes" [8]. Furthermore, the impossibility of [4, 12] says that the overall decision-making process cannot possibly equalize all four measures. However, in some cases one can provide alternative (non-statistical) justification for the fairness of the overall process: For instance, if the downstream decision process never errs, the overall process

might be considered "procedurally fair." We present more detailed reflections on our deferral-based approach in Section 6.

We note that deferring was considered in machine learning in a number of contexts, including the context of fairness-preservation [13]. In these works, the classifier typically defers only when its confidence regarding some decision is low. By contrast, we use deferrals in order to "equalize" the probability mass functions of the soft classifier over the two groups, which may involve deferring on individuals for whom there is higher confidence. Furthermore, our framework allows for a wide range of deferral strategies which might be used to promote additional goals. Pursuing alternate strategies for deferral is an interesting direction for future work.

Experimental results. We demonstrate the validity of our methodology on the Broward county dataset with COMPAS scores made public by ProPublica [1]. Indeed, it has been shown that the COMPAS scoring mechanism is an approximately calibrated soft classifier. We first ran our two-threshold post-processing mechanism and obtained a binary decision algorithm which equalizes both PPV and NPV across Caucasians and African-Americans.

We then ran our post-processing mechanism with deferrals to equalize all four of PPV, NPV, FPR, FNR across the two groups, with three different methods for deciding how to defer: In the first method, decisions are deferred only for Caucasians; in the second, decisions are deferred only for African Americans; in the third method, decisions are deferred for an equal fraction of Caucasians and of African Americans. This fraction is precisely equal to the statistical (total variation) distance between the profiles of scores produced by the soft classifier on the two groups. More details about the results are given in Section 5.

1.1 Related work

We briefly describe the works most closely related to ours, though both the list of works and their summaries are inevitably too short. Our work fits in a research program on group fairness notions following the work of Chouldechova [4] and Kleinberg et al. [12]. Our work considers the notions of calibration as formalized in [16] and those of PPV, NPV, FPR, and FNR from [4] and [12].

The power of post-processing calibrated scores into decisions using threshold classifiers in the context of fairness has been previously studied by Corbett-Davies, Pierson, Feller, Goel, and Huq [6]. As in our work, they show that it is feasible to equalize certain statistical fairness notions across groups using (possibly different) thresholds. They additionally show that these thresholds are in some sense optimal. Whereas [6] focuses on statistical parity, conditional statistical parity, and false positive rate, our most comparable results consider PPV. In our work, we further show that in some cases thresholds fail to equalize both PPV and NPV (called *predictive parity* by [4]), unless we also allow our post-processor to defer on some inputs. Our work also studies methods of post-processing that are much more powerful than thresholding, especially when allowing deferrals.

Using deferrals to promote fairness was also considered by Zemel, Madras, and Pitassi [13]. Specifically they consider how deferring on some inputs may promote a combination of accuracy and fairness, especially when taking explicit account of the downstream decision maker. They make use of two-threshold deferring post-processors like those discussed in Section 4. [13] takes a more experimental approach and focuses on minimizing the "disparate impact," a measure of total difference in classification error between groups, while maximizing accuracy. One important difference between our works is that Madras et al. distinguish between "rejecting" and "deferring." Rejecting is oblivious as to properties of the downstream decision maker, while deferring tries to counteract the biases of the decision maker. Our work considers only the former notion, but uses the term "defer" instead of "reject."

Our work inherits both the strengths and weakness of the group fairness research program. The clear definitions and goals of group fairness have have catalyzed the field and caused rapid progress: early infeasibility results [4, 12], positive results for complex and intersecting collections of groups [10, 11], and extensions to the basic model—including [13, 14] and this work. The formalization of group fairness has fostered precise discussion and greater understanding, including of its shortcomings. Group fairness notions have been criticized for not fully capturing the complex social goals that motivate our community's interest in fairness: failure to compose [8], failure to adequately capture people's wellbeing [5], and failure to prevent against certain social evils [7]. However, we are optimistic that improving our understanding of group fairness will contribute to the successful study of algorithmic fairness generally.

2 PRELIMINARIES

We study the problem of binary classification. An *instance* is an element, usually denoted x, of a universe $\mathcal{X}$. We restrict our attention to instances sampled uniformly at random from the universe, denoted $X \sim \mathcal{X}$. Our theory extends directly to any other distribution on $\mathcal{X}$; that distribution does not need to be known to the classifiers. Each instance x is associated with a *true type* $Y(x) \in \{0,1\}$. Each instance x is also associated with a *group* $G(x) \in \mathcal{G}$, where $\mathcal{G}$ is the set of groups. We restrict our attention to sets $\mathcal{G}$ that form a partition of the universe $\mathcal{X}$. We denote by $\mathcal{X}_g$ the set of instances x in group g, and by X_g the random variable distributed uniformly over $\mathcal{X}_g$. Note that for any events E_1 and E_2, $\Pr_{X\sim\mathcal{X}_g}[E_1 \mid E_2] = \Pr_{X\sim\mathcal{X}}[E_1 \mid E_2, G(X) = g]$.

Definition 2.1 (Base rate (BR))**.** The *base rate* of a group $g \in \mathcal{G}$, is

$$\mathrm{BR}_g = \Pr[Y(X_g) = 1] = \mathbb{E}[Y(X_g)].$$

When $\mathcal{X}$ is finite, BR_g is simply the fraction of individuals x in the group g for whom $Y(x) = 1$.

A *classifier* is a randomized function with domain $\mathcal{X} \times \mathcal{G}$. A *hard classifier*, denoted $\hat{Y}$, outputs a *prediction* in $\{0,1\}$, interpreted as a guess of the true type $Y(x)$. A *soft classifier*, denoted $\hat{S}$, outputs a *score* $s \in [0,1]$, interpreted as a measure of confidence that $Y(x) = 1$. We restrict our attention to soft classifiers with finite image. We call a classifier *group blind* if its output is independent of the input group g. For all groups $g \in \mathcal{G}$, we call a hard classifier $\hat{Y}$ *non-trivial on* g if $\Pr[\hat{Y}(X_g) = 1] > 0$ and $\Pr[\hat{Y}(X_g) = 0] > 0$. Hard classifiers are *trivial on* g if they are not non-trivial on g.

A *post-processor* is a randomized function with domain $[0,1] \times \mathcal{G}$. As with classifiers, a post-processor can be *hard* or *soft*. A hard post-processor, denoted $\hat{D}$, outputs a prediction in $\{0,1\}$. A soft post-processor, denoted $\hat{D}^{\mathrm{soft}}$, outputs a score $s \in [0,1]$. Observe that for a soft classifier $\hat{S}$, $\hat{D} \circ \hat{S}$ is a hard classifier, and $\hat{D}^{\mathrm{soft}} \circ \hat{S}$ is

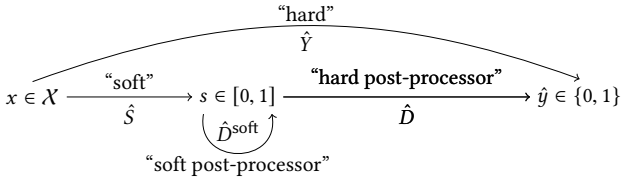

Figure 1: We call a classifier that returns results in $[0,1]$ ***a* soft classifier *to differentiate it from those which return results in*** $\{0,1\}$***, which we call* hard classifiers. *We refer to classifiers that take as input the output of a soft classifier as* post-processors.

a soft classifier. As with classifiers, we call a post-processor group blind if its output is independent of the group g, and we restrict our attention to post-processors with finite image. The restriction to finite image is for mathematical convenience and also because digital memory leads to discrete universes; our results generalize to infinite images as well.

In Section 4, we expand the definitions of both classifier and post-processors to allow an additional input or output: the special symbol $\perp$, indicating a *deferral*.

2.1 Calibration

Several works concerning algorithmic fairness focus on various notions of *calibration*. The following calibration notions are defined only over soft classifiers:

Definition 2.2 (Calibration (Soft))**.** We say a soft classifier $\hat{S}$ is *calibrated* if $\forall s \in [0,1]$ for which $\Pr_{X\sim\mathcal{X}}[\hat{S}(X) = s] > 0$,

$$\Pr_{X\sim\mathcal{X}}[Y(X) = 1 \mid \hat{S}(X) = s] = \mathop{\mathbb{E}}_{X\sim\mathcal{X}}[Y(X) \mid \hat{S}(X) = s] = s.$$

The probability above is taken over the sampling of X, as well as random choices made by $\hat{S}$ at classification time.

Definition 2.3 (Groupwise Calibration (Soft))**.** We say that a soft classifier $\hat{S}$ is *groupwise calibrated* if it is calibrated within all groups. That is, $\forall g \in \mathcal{G}$ and $\forall s \in [0,1]$ for which $\Pr[\hat{S}(X_g) = s] > 0$, we have that

$$\Pr[Y(X_g) = 1 \mid \hat{S}(X_g) = s] = s.$$

Groupwise calibration is essentially the same notion as *multicalibration* [10] with the difference that in their case the true types are values in $[0,1]$. We use a different term to emphasize that we restrict our attention to collections of groups $\mathcal{G}$ that form a partition of the universe $\mathcal{X}$.

The two definitions above are stated for soft classifiers whose output distribution is discrete, since we must be able to condition on the event $\hat{S}(X) = s$ or $\hat{S}(X_g) = s$. That said, it extends naturally to classifiers with continuously-distributed outputs provided that the conditional probabilities are well defined.

2.2 Accuracy Profiles

Throughout this work, we make repeated reference to the probability mass function of the random variable $\hat{S}(X_g)$ for a calibrated soft classifier $\hat{S}$ acting on a randomly distributed input X_g. We call this distribution on calibrated scores an *accuracy profile* (AP).

Definition 2.4 (Accuracy Profile (AP))**.** The *accuracy profile (AP)* of a calibrated soft classifier $\hat{S}$ for a group g, denoted by $\hat{\mathcal{P}}_g$, is the PMF of $\hat{S}(X_g)$. That is, for $s \in [0,1]$, $\hat{\mathcal{P}}_g(s) = \Pr[\hat{S}(X_g) = s]$.

Abusing notation, we denote by $\hat{\mathcal{P}}$ the collection $\{\hat{\mathcal{P}}_g\}_{g\in\mathcal{G}}$, and call it the *AP of* $\hat{S}$. We denote by $\text{Supp}(\hat{\mathcal{P}}_g)$ the support of the AP $\hat{\mathcal{P}}_g$, namely the set $\text{Supp}(\hat{\mathcal{P}}) = \{s : \exists x \in \mathcal{X}_g, \exists r \text{ s.t. } \hat{S}(x,r) = s\} \subseteq [0,1]$.

An accuracy profile is a distribution of scores for a calibrated classifier $\hat{S}$. Because $\hat{S}$ is calibrated, the AP conveys information about the performance of $\hat{S}$, and is constrained by properties of the underlying distribution on X. For example, the AP's expectation is exactly the base rate for the population:

Proposition 2.1 (Proof in [3]). *For any groupwise calibrated soft classifier $\hat{S}$, for all groups $g \in \mathcal{G}$: $BR_g = \mathbb{E}[\hat{S}(X_g)]$.*

Accuracy profiles also provide useful geometric intuition for reasoning about the effects of post-processing calibrated scores. We elaborate on this in Section 3.1 (see Figure 2).

2.3 Group Fairness Measures

Several well-studied measures of statistical "fairness" (e.g., [4, 9–12, 16]) look at how the following key performance measures of a classifier differ across groups. We define these statistics formally:

Definition 2.5. For a hard classifier $\hat{Y}$ and group g, we define the
false positive rate of $\hat{Y}$ for g:
$$\text{FPR}_{\hat{Y},g} = \Pr[\hat{Y}(X_g) = 1 \mid Y(X_g) = 0];$$
false negative rate of $\hat{Y}$ for g:
$$\text{FNR}_{\hat{Y},g} = \Pr[\hat{Y}(X_g) = 0 \mid Y(X_g) = 1];$$
positive predictive value of $\hat{Y}$ for g:
$$\text{PPV}_{\hat{Y},g} = \Pr[Y(X_g) = 1 \mid \hat{Y}(X_g) = 1];$$
negative predictive value of $\hat{Y}$ for g:
$$\text{NPV}_{\hat{Y},g} = \Pr[Y(X_g) = 0 \mid \hat{Y}(X_g) = 0].$$

The probability statements in the definitions above reflect two sources of randomness: the sampling of X_g from the group g and any random choices made by the classifier $\hat{Y}$.

Among previous works, some [9, 12, 15] focus on equalizing only one or both of the false positive rates and false negative rates across groups, called *balance* for the negative and positive classes, respectively. Equalizing positive and negative predictive value across groups is often combined into one condition called *predictive parity* [4]. We split the value out to be a separate condition for the positive and negative predictive classes. Predictive parity appears to be a hard-classifier analogue of calibration: both can be interpreted as saying that the output of the classifier (hard or soft) contains all the information contained in group membership. Our results highlight that the relationship between these notions is more subtle than it first appears; see Section 3 for further discussion.

3 THE LIMITS OF POST-PROCESSING

Suppose throughout this section that $\hat{S}$ is a groupwise calibrated soft classifier. Our goal in this section is to make binary predictions based on $\hat{S}(x)$ — and possibly the group $G(x)$ — subject to equalizing PPV among groups. That is, we wish to make a prediction using a hard post-processor $\hat{D}$ such that $\hat{Y} = \hat{D} \circ \hat{S}$ equalizes PPV among groups. We chose to concentrate first on (the limitations of) equalizing PPV rather than FPR and FNR due to the conceptual similarity of PPV to calibration, and we address NPV in [3]. Also, the case of equalizing false positive rates with thresholds is addressed in [6].

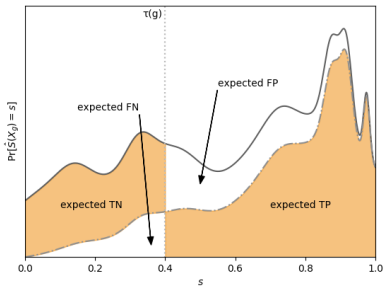

Figure 2: Accuracy profiles (APs, definition 2.4) yield useful geometric intuitions, which come from the calibration property (definition 2.2). The dashed line is the AP multiplied by the $y = x$ line; the region below this line shows the expected positives and above shows the expected negatives. With a threshold, the expected PPV, NPV, FPR, and FNR can be seen visually.

3.1 Fairness Conditions for Post-Processors

We begin by making a simple observation about post-processing that provides some geometric intuition for the rest of this section. Just as in Proposition 2.1, we can express $\text{PPV}_{\hat{Y},g}$ succinctly in terms of conditional expectations over the AP $\hat{\mathcal{P}}_g$. We state this formally in Proposition 3.1.

Proposition 3.1 (Proof in [3]). *Let $\hat{Y} = \hat{D} \circ \hat{S}$ be a hard classifier that is non-trivial for all $g \in \mathcal{G}$ where $\hat{S}$ is groupwise calibrated with respect to $\mathcal{G}$. For any $g \in \mathcal{G}$ we have:*

$$PPV_{\hat{Y},g} = \mathbb{E}[\hat{S}(X_g) \mid \hat{Y}(X_g) = 1]$$

This characterization of PPV and NPV in terms of conditional expectations lets us geometrically see how certain post-processing decision rules will interact with the AP for a group g. For example, Figure 2 shows the expected true positives, true negatives, false positive, and false negatives when using a threshold.

3.2 General impossibility of equalizing PPV

It is not always possible to directly post-process a soft groupwise calibrated classifier into a hard one with equalized PPV (or NPV) for all groups, as we demonstrate by counterexample in Proposition 3.2. Before proceeding, we note that our counterexample is somewhat contrived—in particular, the AP induced by the soft classifier $\hat{S}$ in the proof of Proposition 3.2 takes only one value on each group. When the AP of $\hat{S}$ is more nicely structured on each group, we will see that there are general methods to equalize PPV (or NPV).

Proposition 3.2. *Fix two disjoint groups g_1 and g_2 with respective base rates BR_1 and BR_2 such that $BR_1 \neq BR_2$. Then there exists a soft classifier $\hat{S}$ that is groupwise calibrated, but for which there is no post-processor $\hat{D} : [0,1] \times \mathcal{G} \to \{0,1\}$ such that $\hat{D} \circ \hat{S}$ equalizes PPV, unless $\Pr[\hat{D}(BR_i, g_i) = 1] = 0$ for $i = 1$ or 2.*

Proof of Proposition 3.2. Consider the classifier $\hat{S}$ such that $\hat{S}(x) = \text{BR}_1$ if $x \in g_1$ and $\hat{S}(x) = \text{BR}_2$ if $x \in g_2$. This classifier is trivially groupwise calibrated. Since $\Pr[\hat{D}(\text{BR}_i, g_i) = 1] > 0$ for $i = 1$ and 2, we conclude that $\text{PPV}_{\hat{Y},g_i}$ is well-defined for g_1 and g_2. The proof now follows from the characterization of PPV in Proposition 3.1. This is because $\text{PPV}_{\hat{Y},g_i}$ is equal to the expectation of $\hat{S}(X)$ where X is drawn from a distribution with support contained in g_i, and hence it is equal to BR_i, and $\text{BR}_1 \neq \text{BR}_2$. □

3.2.1 *A Niceness Condition for APs.*

Motivated by the above proof, we define a non-degeneracy condition on APs.

Definition 3.1 (Niceness of APs)**.** Let $\mathcal{G}$ be a set of groups. An accuracy profile $\hat{\mathcal{P}}$ is *nice* if $\text{Supp}(\hat{\mathcal{P}}_g)$ is the same for all $g \in \mathcal{G}$.

Note that this condition rules out the counterexample given by Proposition 3.2, since the AP in the counterexample had different (in fact, disjoint) supports for different groups.

3.3 Equalizing PPV or NPV by Thresholding

We pay special attention to thresholds because they are simple to understand and therefore very widely used. We use one slight modification to deterministic thresholds that adds an element of randomness: if a score is *at* the threshold, we randomly determine which side of the threshold it falls on, according to a distribution defined below.

Definition 3.2 (Threshold Post-Processor)**.** A threshold post-processor $\hat{D}_{(\tau,\mathcal{R})} : [0,1] \times \mathcal{G} \to \{0,1\}$ is a function from a score $s \in [0,1]$ and a group $g \in \mathcal{G}$, parameterized by τ and $\mathcal{R}$. The threshold parameter $\tau : \mathcal{G} \to [0,1]$ specifies the threshold for the group g, and $\mathcal{R} : \mathcal{G} \to [0,1]$ is the probability of returning 1 when the input score s is on the threshold $\tau(g)$. It returns the following outputs:

$$\hat{D}_{(\tau,\mathcal{R})}(s,g) = \begin{cases} 1 & s > \tau(g) \\ 0 & s < \tau(g) \\ 1 \text{ w.p. } \mathcal{R}(g) \text{ else } 0 & s = \tau(g) \end{cases}$$

In the setting of an infinite number of scores and a continuous domain (i.e. scores are represented by a probability density function instead of a probability mass function), we can use purely deterministic threshold functions in which $\mathcal{R} \equiv 1$, and achieve very similar results for the rest of this section.

We now study the effectiveness of thresholds for post-processing soft classifiers with nice APs. The main takeaways are:

(1) If the AP is nice, threshold post-processors can equalize PPV.
(2) However, group blind threshold post-processors are rather limited in their ability to equalize PPV.
(3) Furthermore, equalizing PPV with thresholds (group blind or otherwise) may have undesirable social consequences (see Figure 3).
(4) Thresholds cannot always equalize PPV and NPV simultaneously, even for nice APs (Proposition 3.3).

Results 1-3 also apply to NPV. We delegate formal statements and proofs for Results 1-3 to the full version [3]. Result 4 shows that threshold post-processors are inherently limited, even when they factor in groups.

Proposition 3.3 (Proof in [3]). *Fix groups g_1 and g_2. There exists a soft classifier $\hat{S}$ with a nice AP $\hat{\mathcal{P}}$ such that no threshold post-processor can simultaneously equalize PPV and NPV between groups g_1 and g_2.*

3.4 Equalizing Accuracy Profiles

While thresholding is a conceptually simple approach to post-processing a soft classifier, its power is limited. We now consider a very different approach using soft post-processors to equalize the APs across groups of a soft classifier. The intuition is that if the APs

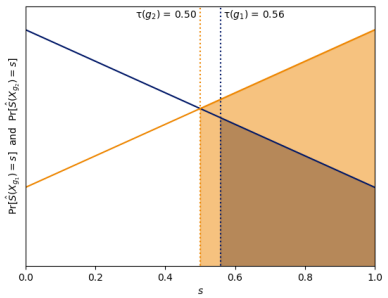

Figure 3: The PPV for both groups is 0.77. However, the threshold for g_1 (dark blue) is higher than the threshold for g_2 (orange), even though g_2 has a higher base rate.

are equal across groups, then any hard post-processor that is *group blind* should result in equal PPV, NPV, FPR, and FNR. We formalize this intuition in Claim 3.1.

Let $\hat{S}$ be a soft classifier and for each group $g \in \mathcal{G}$, let $\hat{\mathcal{P}}_g$ be the AP of $\hat{S}$ for group g. For a soft post-processor $\hat{D}^{\text{soft}}$, let $\hat{S}' = \hat{D}^{\text{soft}} \circ \hat{S}$ and let $\hat{\mathcal{P}}'_g$ be the corresponding AP for group g.

Our goal is to find a soft post-processor $\hat{D}^{\text{soft}}$ such that $\hat{S}'$ is groupwise calibrated, and $\hat{\mathcal{P}}'_g = \hat{\mathcal{P}}'_{g'}$ for all $g, g' \in \mathcal{G}$. In this section, we describe only one approach to constructing $\hat{D}^{\text{soft}}$ which we call *mass averaging*.

The approach of equalizing APs has a fundamental weakness: if $\hat{\mathcal{P}}'_g = \hat{\mathcal{P}}'_{g'}$ and both are calibrated, then $\text{BR}_g = \text{BR}_{g'}$. This severely limits applicability of this approach. However, this limitation will removed in Section 4.2 by allowing deferrals.

Claim 3.1 (Proof in [3]). *If the APs are equal for two groups, then PPV, NPV, FPR, and FNR are equalized by any hard post-processor $\hat{D}$ satisfying group blindness.*

The group-blindness requirement in the claim is necessary: consider the (not group blind) post-processor that outputs 0 on one group and 1 on the other; PPV will not be equalized.

3.4.1 Mass Averaging. The mass-averaging technique is best illustrated with an example. Suppose that $\hat{\mathcal{P}}_{g_1}$ is uniform over $\{0, 0.5, 1\}$, and $\hat{\mathcal{P}}_{g_2}$ is uniform over $\{0, 1\}$. It is easy to define a soft post-processor $\hat{D}^{\text{soft}}$ which equalizes these two APs. On g_1, we leave the score unchanged: $\hat{D}^{\text{soft}}(s, g_1) = s$. On g_2, we compute the output as

$$\hat{D}^{\text{soft}}(s, g_2) = \begin{cases} s & \text{w.p. } 2/3 \\ 0.5 & \text{w.p. } 1/3 \end{cases}.$$

The APs for groups g_1 and g_2 of the resulting soft classifier $\hat{S}' = \hat{D}^{\text{soft}} \circ \hat{S}$ are equal, and are equal to $\hat{\mathcal{P}}_{g_1}$.

In the example, the probability mass is being redistributed by averaging the scores. This can be equivalently viewed as adding noise to the scores and then recalibrating the scores, something discussed in [6].

More generally, a mass-averaging post processor $\hat{D}^{\text{soft}}$ assigns to each possible pair (s, g) a distribution over possible output scores s'. Such a $\hat{D}^{\text{soft}}$ is fully specified by $k \cdot k' \cdot |\mathcal{G}|$ parameters, where k is the number of possible values of s and k' is the number of possible values of s'. Given a soft classifier $\hat{S}$ and a mass-averaging post processor $\hat{D}^{\text{soft}}$, the constraint that the resulting APs are equalized across groups is linear in these parameters. Such classifiers, therefore, may be found by a linear program. We do not explore the choice of mass-averaging post-processors further.

4 POST-PROCESSING CALIBRATED CLASSIFIERS WITH DEFERRALS

In the first part of the paper, we considered the problem of post-processing calibrated soft classifiers, which output a score $s \in [0, 1]$, into fair hard classifiers, which output a decision in $\hat{y} \in \{0, 1\}$, subject to a number of group fairness conditions. In the remainder of this work, we reconsider this problem, but with one important change: we allow classifiers to "refuse to decide" by outputting the special symbol $\perp$. We call such classifiers *deferring* classifiers, borrowing the nomenclature from [13]. The output $\perp$ is the deferring classifier's way of refusing to make a decision and deferring to a downstream decision maker. For example, a risk assessment tool might aid a parole board to make a decision by categorizing an individual as high risk or low risk, or it might output $\perp$—providing no advice and deferring to the judgment of the board.

We now modify our notation appropriately. Instances x are still associated with a true type $Y(x) \in \{0, 1\}$ and a group $G(x) \in \mathcal{G}$. A deferring hard classifier $\hat{Y}$ is a randomized function $\hat{Y} : \mathcal{X} \to \{0, 1, \perp\}$. A deferring soft classifier is a randomized function $\hat{S} : \mathcal{X} \to [0, 1] \cup \{\perp\}$. A deferring hard (resp. soft) post-processor is a randomized function $\hat{D} : [0, 1] \cup \{\perp\} \times \mathcal{G} \to \{0, 1, \perp\}$ (resp. $\hat{D}^{\text{soft}} : [0, 1] \cup \{\perp\} \times \mathcal{G} \to [0, 1] \cup \{\perp\}$) that takes as input the output of a deferring soft and post-processes it into a deferring hard (resp. soft) classifier. We also introduce new versions of the FPR and FNR, conditioned on not deferring.

Definition 4.1. The *conditional false positive rate* and *conditional false negative rate* of a deferring hard classifier $\hat{Y}$ for a group g are, respectively:

$$\text{cFPR}_{\hat{Y},g} = \Pr[\hat{Y}(X_g) = 1 \mid Y(X_g) = 0, \hat{Y}(X_g) \neq \perp]$$
$$\text{cFNR}_{\hat{Y},g} = \Pr[\hat{Y}(X_g) = 0 \mid Y(X_g) = 1, \hat{Y}(X_g) \neq \perp].$$

We additionally consider a version of the accuracy profile conditioned on not deferring, which we call the *conditional AP*. For non-deferring soft classifiers, Definitions 4.2 and 2.4 coincide.

Definition 4.2. The *conditional AP* $\hat{\mathcal{P}}_g$ of a classifier $\hat{S}$ for a group g is the PMF of $\hat{S}(X_g)$, conditioned on not outputting $\perp$. That is, for $s \in [0, 1]$, $\hat{\mathcal{P}}_g(s) = \Pr[\hat{S}(X_g) = s \mid \hat{S}(X_g) \neq \perp]$. Note that the conditional AP is undefined if $\Pr[\hat{S}(X_g) \neq \perp] = 0$.

Abusing notation, we denote by $\hat{\mathcal{P}}$ the collection $\{\hat{\mathcal{P}}_g\}_{g \in \mathcal{G}}$, and call it the *conditional AP of* $\hat{S}$.

The conditional error rates are applicable generally, but they can be difficult to interpret. The consequences of using the conditional FPR and FNR are discussed further in Section 6 along with a discussion of different deferral models. They are also amenable to the consideration of additional goals which we will briefly address. For example, one could seek to minimize the total deferral rate, equalize the deferral rate among groups, or prefer deferrals on positive instances.

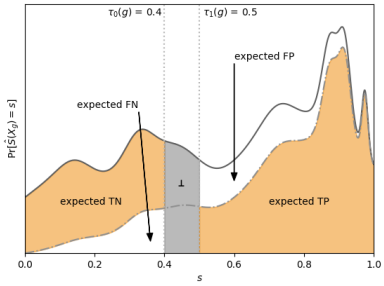

Figure 4: Threshold post-processors with deferrals defer between the thresholds.

4.1 Thresholding with deferrals

We return now to the problem of post-processing of calibrated soft classifiers, but now with the extra power of deferring on some inputs. We revisit the two approaches discussed in Section 3: thresholding and equalizing accuracy profiles.

Proposition 3.2 stated PPV and NPV cannot both be equalized across groups in general when using only a single threshold per group. By using two thresholds per groups and deferring on some inputs, PPV and NPV can always be equalized across groups.

We post-process using two thresholds per group as follows: return 0 when s is lower than the first threshold, return $\perp$ between the thresholds, and return 1 above the second threshold, as shown in Figure 4. This buys us more degrees of freedom when equalizing binary constraints, and it has the useful property that we say $\perp$ on the instances where we are the least confident about the predicted type. We adapt our notation as follows:

Definition 4.3 (Deferring Threshold Post-Processor)**.** A deferring threshold post-processor $\hat{D}_{(\tau_0,\tau_1,\mathcal{R}_0,\mathcal{R}_1)}$ assigns to each group g two thresholds $\tau_0(g), \tau_1(g) \in \text{Supp}(\hat{\mathcal{P}}_g)$, and two probabilities $\mathcal{R}_0(g), \mathcal{R}_1(g) \in [0,1]$, with the following requirements:

(1) for all $g \in \mathcal{G}$, $\tau_0(g) \leq \tau_1(g)$
(2) for all $g \in \mathcal{G}$ for which $\tau_0(g) = \tau_1(g)$, $\mathcal{R}_1(g) + \mathcal{R}_0(g) \leq 1$. This corresponds to the case where the two thresholds are the same, and therefore individuals with that score must be mapped to 1 with probability $\mathcal{R}_1(g)$, and to 0 with probability $\mathcal{R}_0(g)$, with the remainder mapped to $\perp$.

The corresponding threshold post-processor is defined as follows:

$$\hat{D}_{(\tau_0,\tau_1,\mathcal{R}_0,\mathcal{R}_1)}(s,g) = \begin{cases} 1 & s > \tau_1(g) \\ 0 & s < \tau_0(g) \\ \perp & \tau_0(g) < s < \tau_1(g) \\ 1 \text{ w.p. } \mathcal{R}_1(g), \text{ else } \perp & s = \tau_1(g) \\ 0 \text{ w.p. } \mathcal{R}_0(g), \text{ else } \perp & s = \tau_0(g) \\ 1 \text{ w.p. } \mathcal{R}_1(g), 0 \text{ w.p. } \mathcal{R}_0(g), \text{ else } \perp & s = \tau_0(g) = \tau_1(g) \\ \perp & s = \perp \end{cases}$$

Using two thresholds allows the equalization of both PPV and NPV across groups in general, whereas without deferrals we could only equalize one or the other. In our full paper [3], we first demonstrate the existence of near-trivial classifiers that equalize PPV and NPV by deferring on all but the highest and lowest scores. We now claim the existence of meaningful non-trivial threshold deferring post-processors that equalize PPV and NPV across groups.

Proposition 4.1 (Proof in [3]). *Let $\hat{S}$ be a soft classifier with nice AP that is groupwise calibrated for a set of groups $\mathcal{G}$. Suppose that $|\text{Supp}(\hat{\mathcal{P}}_g)| \geq 2$ for all $g \in \mathcal{G}$. Then there exists a non-trivial threshold post-processor $\hat{D}_{(\tau_0,\tau_1,\mathcal{R}_0,\mathcal{R}_1)}$ such that the hard classifier $\hat{Y} = \hat{D}_{(\tau_0,\tau_1,\mathcal{R}_0,\mathcal{R}_1)} \circ \hat{S}$ equalizes PPV_g and NPV_g for all $g \in \mathcal{G}$.*

The following example demonstrates that it is sometimes possible to equalize PPV, NPV, FPR, and FNR using deferrals, but without equalizing the APs themselves:

Example 4.1 (Equalizing PPV, NPV, cFPR, and cFNR with Thresholds). *This example is presented with continuous support $[0,1]$ for simplicity. Consider two APs, one for group g_1 and one for g_2. Let the AP for g_1 be uniform (with density give by the line $\hat{\mathcal{P}}(s) = 1$), and let the AP for group g_2 have density given by the parabola $\hat{\mathcal{P}}(s) = 6s(1-s)$, as shown in Figure 5.*

Consider the post-processor $\hat{D}^{\text{soft}}_{(\tau_0,\tau_1)}$. Let $\tau_0(g_1) = \tau_0(g_1) = 0.5$, let $\tau_0(g_2) = \frac{1}{6}(5-\sqrt{7})$ and let $\tau_1(g_2) = 1 - \frac{1}{6}(5-\sqrt{7})$ as shown in Figure 5.

The PPV and NPV of both groups is $\frac{3}{4}$, and the cFPR and cFNR of both is $\frac{1}{4}$, thus equalizing all four values.

This example is somewhat unsatisfactory because the base rates are equal in the two groups. We did not find a similar example without equal base rates.

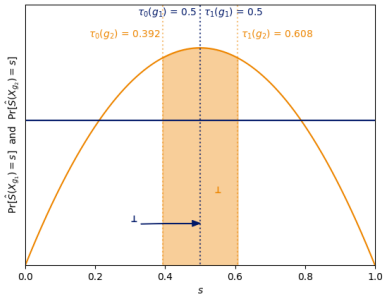

Figure 5: This threshold post-processor equalizes PPV, NPV, cFPR, and cFNR as described in Example 4.1.

4.2 Equalizing APs with deferrals

As with Claim 3.1, equalizing the conditional APs between groups renders trivial the task of downstream decision-making subject to equality of PPV, NPV, cFPR, and cFNR. Importantly, unlike in Section 3.4, equalizing the conditional APs between groups does not require the groups to have equal base rates, greatly increasing the applicability of this approach.

Claim 4.1. *If the conditional APs are equal for two groups, then PPV, NPV, cFPR, and cFNR are equalized (or simultaneously undefined) by any hard deferring post-processor $\hat{D}$ satisfying (1)* group blindness *and (2) $\hat{D}(\perp, g) = \perp$ $(\forall g)$.*

The additional condition—that $\hat{D}$ defers on input $\perp$—is necessary: if $\hat{D}$ output 1 on all inputs (even on $\perp$), then PPV would remain unequal as long as the base rates differed. The proof is similar to the proof of Claim 3.1 and is included in the full version [3].

Deferrals are a powerful tool for manipulating, and thereby equalizing, conditional APs. Consider a function $Q : (s,g) \mapsto [0,1]$

$$\hat{D}^{\text{soft}}_{Q}(s,g) = \begin{cases} \perp & \text{if } s = \perp \\ \perp \text{ w.p. } Q(s,g), \text{ else } s & \text{otherwise} \end{cases}$$

If $\hat{S}$ is a calibrated classifier, the soft deferring classifier $\hat{S}' := \hat{D}_Q^{\text{soft}} \circ \hat{S}$ is still calibrated. For a group g, let $\hat{\mathcal{P}}_g$ be the AP of $\hat{S}$ and $\hat{\mathcal{P}}'_g$ be the AP of $\hat{S}'$. There is a simple graphical intuition for the shape of $\hat{\mathcal{P}}'_g$, as shown in Figure 6.

The following theorem, proved in the full version [3], states that any conditional AP can be transformed into almost any other conditional AP by appropriate choice of Q.

Theorem 4.1. *Let $\hat{\mathcal{P}}_g$ be a conditional AP of a soft classifier $\hat{S}$ on group g, and let $\hat{\mathcal{P}}^*$ be any probability mass function such that $\text{Supp}(\hat{\mathcal{P}}^*) \subseteq \text{Supp}(\hat{\mathcal{P}}_g)$. Then there exists Q for which the calibrated AP $\hat{\mathcal{P}}'_g$ of $\hat{D}_Q^{\text{soft}} \circ \hat{S}$ is equal to $\hat{\mathcal{P}}^*$.*

Together, Theorem 4.1 and Claim 4.1 suggest a general framework for using deferrals to post-process a soft, possibly deferring classifier $\hat{S}$ which is groupwise calibrated into a hard deferring classifier which simultaneously equalizes PPV, NPV, cFPR, and cFNR across groups, as follows.

For each $g \in \mathcal{G}$, let $\hat{\mathcal{P}}_g$ be the conditional AP of $\hat{S}$ for group g. Let $\hat{\mathcal{P}}^*$ be any conditional AP such that $\text{Supp}(\hat{\mathcal{P}}^*) \subseteq \cap_{g \in \mathcal{G}} \text{Supp}(\hat{\mathcal{P}}_g)$. Use Theorem 4.1 to equalize the conditional AP for all groups $g \in \mathcal{G}$. Then use any hard post-processor $\hat{D}$ satisfying the requirements of Claim 4.1 to make the ultimate deferring hard classifier. This method is shown in Figure 6.

This framework allows for enormous flexibility in the choice of both $\hat{\mathcal{P}}^*$ and $\hat{D}$, even when considering just two groups g_1 and g_2. In Figure 9, we illustrate the first step of the framework on a COMPAS dataset using $\min\{\hat{\mathcal{P}}_{g_1}, \hat{\mathcal{P}}_{g_2}\}$ as $\hat{\mathcal{P}}^*$, where g_1 is African-Americans and g_2 is Caucasians. In Figure 8 in Section 5, we also use $\hat{\mathcal{P}}_{g_1}$ and $\hat{\mathcal{P}}_{g_2}$ as $\hat{\mathcal{P}}^*$.

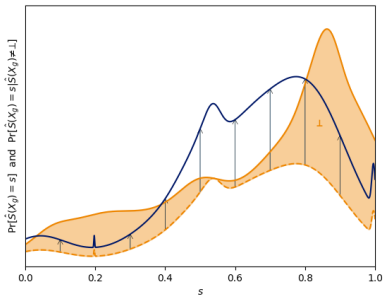

Figure 6: Choosing deferrals appropriately allows transforming one AP into another (conditional) AP. In this example, the solid orange line is the original AP $\hat{\mathcal{P}}_g = \Pr[\hat{S}(X_g) = s]$. By deferring at the rates indicated by the shaded region, the resulting conditional AP $\hat{\mathcal{P}}'_g = \Pr[\hat{S}(X_g) = s \mid \hat{S}(X_g) \neq \perp]$ is represented by the dark blue line. The area of the shaded region is Δ.

One can design $\hat{\mathcal{P}}^*$ to achieve additional goals. For example, the choice $\hat{\mathcal{P}}^* = \min\{\hat{\mathcal{P}}_{g_1}, \hat{\mathcal{P}}_{g_2}\}$ results in *equal deferral rate* across each group (equal to the total variation distance between the two initial conditional APs). The framework can be further expanded by combining deferrals with other methods for manipulating conditional APs, including the mass-averaging discussed in Section 3.4. A better understanding of these techniques is left for future work.

5 EXPERIMENTS ON COMPAS DATA

We demonstrate the validity of our methodology on the Broward County data containing the recidivism risk decile scores of the COMPAS tool [1]. We restrict our attention to the subset of defendants whose race is recorded as either African-American or Caucasian. Our code can be found at: https://github.com/nishanthdikkala/postprocessing-deferrals

It has been shown that the COMPAS scoring mechanism is an approximately calibrated soft classifier with 10 possible outcomes across the two race groups of African-Americans and Caucasians. We note here that the distribution of the COMPAS scores differs significantly across the two groups. In particular, the scores for African-Americans are more evenly distributed as opposed to the skewed distribution seen with Caucasians.

Thresholding with Deferrals. We first ran our two-threshold post-processing mechanism (Section 4.1) and obtained a binary decision algorithm with deferrals which equalizes both PPV and NPV across Caucasians and African-Americans (See Figure 7). We observe that the percent of deferrals in total is smaller than 20% of the decisions to be made which shows that a fairly large number of the defendants can be classified in this manner without having to defer to a downstream decision maker.

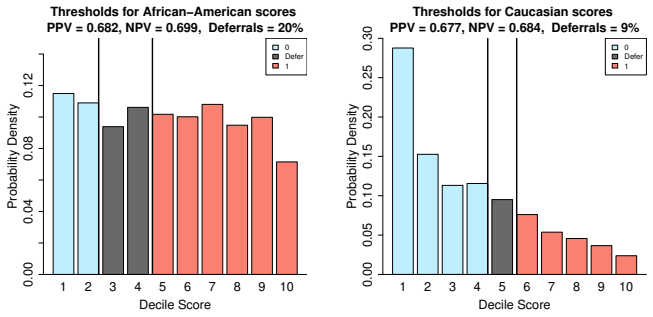

Figure 7: Two thresholds are applied to each AP for the COMPAS data from 2016, (approximately) equalizing PPV and NPV. In the left (right) plot we show the thresholds for the African American (Caucasian) group.

Next we look at our post-processing mechanisms to equalize all four quantities PPV, NPR, FPR, and NPR using deferrals (Section 4.2). As was noted earlier in the paper, equalizing the APs of the two groups post-deferral achieves the goal of equalizing all four of the above quantities. We implement two methods for doing so.

Converting one AP into Another. In the first method, decisions are deferred only on one group so as to convert its AP into that of the other group. First, we consider deferring only on African-Americans to convert their AP into that of Caucasians (left side of Figure 8); next, decisions are deferred only for Caucasians (right side of Figure 8).

Equalizing APs. Alternately we have a second method where decisions are deferred for an equal fraction of Caucasians and of African Americans (Figure 9). This fraction is precisely the *total variation* distance between the two APs.

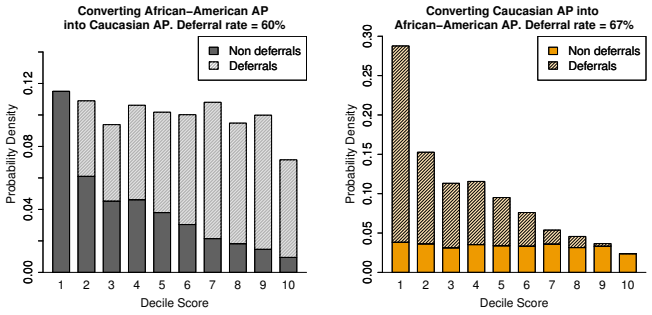

Figure 8: Two instances of our conditional AP equalization method applied to COMPAS data from 2016. On the left (right) plot, we use deferrals to create a conditional AP for African-Americans (Caucasians) that matches the AP for Caucasians (African-Americans).

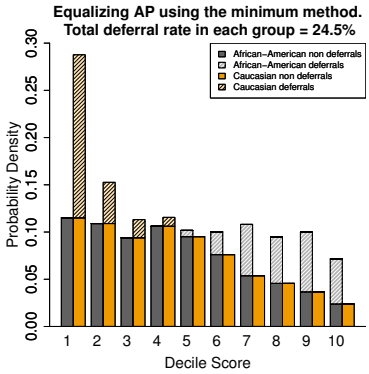

Figure 9: A version of our conditional AP equalization method. The conditional AP has the same distribution as the pointwise minimum of the two APs. The total deferral rate is equalized across the two groups (equals the TV distance between the two APs), but the distribution of deferrals across scores is not.

Observations. We observe several interesting phenomena on the COMPAS data set. First, using the method of deferring only on African-Americans we defer on roughly 36% of the total decisions. This number goes down to roughly 25% when we defer only on Caucasians. This seems to suggest as a general heuristic to try and use deferrals on the group with smaller size. The total deferral fraction is also roughly 25% when we defer on an equal fraction of Caucasians and African-Americans.

Second, for all three methods that equalize the accuracy profiles, for this particular dataset, deferrals happen more on the "extremes", namely on individuals with respect to which the classifier had relatively high confidence (either close to 0 or close to 1). This stands in sharp contrast to how the two-threshold method (Figure 7) distributes its deferrals—they occur in the middle of the distribution (examples on which the classifier is "unsure"). While it may seem somewhat counter-productive to defer on these individuals, any method that seeks to first equalize the accuracy profiles will have to defer most on the scores which appear in different probabilities across the two groups (which, for the COMPAS predictor, is at the extremes). Furthermore, deferring on such scores may make sense from a social point of view: When a score appears at drastically different rates for different groups, perhaps deferring to another decision mechanism can be used to check for systemic bias in the present one.

Alternatively, one can view the above observation as proposing a design criterion for calibrated soft classifiers. That is, if one wants to have a classifier that defers only on individuals for which the classifier has less confidence, and still guarantee that the APs for the protected groups are equal, conditioned on not deferring, then one should design the soft classifier so that the APs for the protected groups are the same (or almost the same) close to 0 and close to 1. Indeed, if the COMPAS classifier had these properties, then our post-processing algorithm would have deferred only (or, mainly) on individuals with low certainty (namely "medium risk" individuals).

6 MODELS OF DEFERRING

Whether or not a classifier is thought of as promoting fairness depends on the context; this is true for both deferring and non-deferring classifiers. In addition to the myriad considerations present for non-deferring classifiers, deferring classifiers and downstream decision makers introduce some additional axes for consideration.

Cost to the individual: Even though it is not intended to be a final decision, a deferral may impose burdensome costs to an individual being classified. It may mean that a defendant remains in jail while additional hearings are scheduled, that invasive and expensive medical tests are ordered, or that continued investigation engenders social stigma. These costs may not be borne equally by all individuals, and may depend on their group membership, their true type, or other factors. For example, a delay in granting a loan to a applicant may overly burden poorer applicants, even those very likely to repay.

Cost to the decider: Allowing deferrals might make the decision process more cost-effective: Given that in most cases making a determination is cheap, one may now invest more in the deferred cases. For instance, a team of trained moderators might be hired to manually review content on which an automated content filter defers, or an expensive investigation might be required to adjudicate insurance claims that are not cut-and-dry.

Accuracy of downstream decision. One reason to defer is to introduce a delay that will allow for a more accurate decision. Thus the usefulness of allowing a classifier to defer depends on the accuracy of the downstream decision maker. Additional medical tests might allow for highly accurate diagnoses. But a judge deciding bail will be prone to a variety of errors and biases.

"Fairness" of downstream decision (and of composed classifier). Similar to the above, the fairness of the downstream decision maker (however one wants to interpret that) will impact our interpretation of the deferring classifier. Here one should take into account also the "procedural" aspect of the two-step evaluation; here it is important that the downstream classifier will be deemed as "more fair" and "more knowledgeable" than the first stage. Exploring fairness criteria for systems of deferring classifiers and downstream decision-makers, e.g. as done in [2] did for non-deferring classifiers, is an interesting direction for future work.

Frequency of decisions. In many settings, the deferring classifier is a fast, automated test (e.g., automated risk assessment) while the downstream decision maker is a slow, manual process (e.g., parole board). However, we anticipate situations in which there may be

repeated deferring classifiers chained together which comprise the complete decision making pipeline. For example, a doctor might have a sequence of diagnostic tests at her disposal as needed, or a bank might allow many rounds of appeal for loan applications, but with lengthy delays. Some applications might even permit hundreds or thousands of near-continuous deferring classifiers. As an example, consider a live video streaming platform that passively monitors streams for inappropriate content in real time. The automated passive monitor might decide the content is inappropriate, and shut it down; appropriate, and continue passive monitoring; or suspicious (by deferring), and begin active monitoring by devoting more computing resources or bringing in a human moderator.

6.1 Technical implications of deferral model

The contextual considerations discussed above directly impact the appropriate application of a deferring classifier and its goals. An obvious goal is to minimize the overall rate of deferrals while maintaining the best possible FPR, FNR, PPV, and NPV for the classifier conditioned on not deferring, and without considering the distribution of deferrals. However, one might desire very different properties from the distribution of deferrals in different contexts. The deferrals may be distributed differently among individuals with different true type, group membership, or soft-classifier scores, while the burden imposed by deferrals and errors may differ greatly between different populations.

In a medical diagnosis scenario, a false negative (i.e., failing to diagnose a disease) may have serious consequences, and deferring to run additional non-invasive and inexpensive additional tests may be generally acceptable. On the other hand, an insurance provider may prefer to minimize expensive investigations by paying out more false claims.

The context may also affect the way one defines the deferral analogues of FPR and FNR. While calibration, PPV, and NPV apply directly to deferring classifiers, it is not clear how best to generalize the definitions of error rates. For example, consider false positive rate: by Definition 2.5, the false positive rate of a non-deferring hard classifier $\hat{Y}$ for a group g is $\mathsf{FPR}_{\hat{Y},g} = \Pr[\hat{Y}(X_g) = 1 \mid Y(X_g) = 0]$.

The approach we take in Section 4 is to condition on not deferring (Definition 4.1). A deferring classifier $\hat{Y}$ that output 1 on half of true negative instances (within a g) would have conditional false positive rate as low as 0.5 (if it never output $\perp$ on true negatives) or as high as 1 (if it never output 0 on true negatives). The conditional false positive rate is agnostic towards the downstream decision maker. It codifies no value judgements as to whether a deferral is desirable or undesirable as an individual nor whether deferrals ultimately result in accurate or inaccurate decisions. This is, itself, a value judgement.

A second approach is to leave the original definition unchanged. The same deferring hard classifier as above would have unconditional false positive rate 0.5. This would be true regardless of whether $\hat{Y}$ output 0 or $\perp$ on the other half of true negative instances. We call this the *unconditional false positive rate*. The unconditional false positive rate effectively categorizes deferrals as correct outputs. This may be appropriate if the downstream decision maker has very high accuracy. If, for example, a doctor orders an additional, more accurate diagnostic test in response to a deferral, the unconditional false positive rate might be appropriate.

Finally, a third approach is to base our measure of inaccuracy on true negatives instead of false positives, a reverse of the above.

Just as in the case of non-deferring classifiers, the relationships among these contrasting group statistics, their meaningfulness in different settings, and their application in different settings are not well understood and deserve further study.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, May 23, 2016.

[2] Amanda Bower, Sarah N Kitchen, Laura Niss, Martin J Strauss, Alexander Vargas, and Suresh Venkatasubramanian. 2017. Fair Pipelines. *arXiv preprint arXiv:1707.00391* (2017).

[3] Ran Canetti, Aloni Cohen, Nishanth Dikkala, Govind Ramnarayan, Sarah Scheffler, and Adam D. Smith. 2018. From Soft Classifiers to Hard Decisions: How fair can we be? *CoRR* abs/1810.02003 (2018). arXiv:1810.02003 http://arxiv.org/abs/1810.02003

[4] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *CoRR* abs/1703.00056 (2017). arXiv:1703.00056 http://arxiv.org/abs/1703.00056

[5] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[6] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. *CoRR* abs/1701.08230 (2017). arXiv:1701.08230 http://arxiv.org/abs/1701.08230

[7] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2011. Fairness Through Awareness. *CoRR* abs/1104.3913 (2011). arXiv:1104.3913 http://arxiv.org/abs/1104.3913

[8] Cynthia Dwork and Christina Ilvento. 2018. Fairness Under Composition. *CoRR* abs/1806.06122 (2018). arXiv:1806.06122 http://arxiv.org/abs/1806.06122

[9] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. 3315–3323. http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning

[10] Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. 2017. Calibration for the (Computationally-Identifiable) Masses. *CoRR* abs/1711.08513 (2017). arXiv:1711.08513 http://arxiv.org/abs/1711.08513

[11] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, Stockholmsmässan, Stockholm Sweden, 2564–2572. http://proceedings.mlr.press/v80/kearns18a.html

[12] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. *CoRR* abs/1609.05807 (2016). arXiv:1609.05807 http://arxiv.org/abs/1609.05807

[13] David Madras, Toniann Pitassi, and Richard S. Zemel. 2017. Predict Responsibly: Increasing Fairness by Learning To Defer. *CoRR* abs/1711.06664 (2017). arXiv:1711.06664 http://arxiv.org/abs/1711.06664

[14] Andrew Morgan and Rafael Pass. 2017. Paradoxes in Fair Computer-Aided Decision Making. *CoRR* abs/1711.11066 (2017). arXiv:1711.11066 http://arxiv.org/abs/1711.11066

[15] Andrew Morgan and Rafael Pass. 2018. Achieving fair treatment in algorithmic classification. In *Theory of Cryptography Conference*. Springer, 597–625.

[16] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On Fairness and Calibration. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 5680–5689. http://papers.nips.cc/paper/7151-on-fairness-and-calibration.pdf