Paradoxes in Fair Computer-Aided Decision Making

Andrew Morgan
Cornell University
asmorgan@cs.cornell.edu

Rafael Pass
Cornell Tech
rafael@cs.cornell.edu

Abstract

Computer-aided decision making—where a human decision-maker is aided by a computational classifier in making a decision—is becoming increasingly prevalent. For instance, judges in at least nine states make use of algorithmic tools meant to determine "recidivism risk scores" for criminal defendants in sentencing, parole, or bail decisions. A subject of much recent debate is whether such algorithmic tools are "fair" in the sense that they do not discriminate against certain groups (e.g., races) of people.

Our main result shows that for "non-trivial" computer-aided decision making, either the classifier must be discriminatory, or a rational decision-maker using the output of the classifier is forced to be discriminatory. We further provide a complete characterization of situations where fair computer-aided decision making is possible.

As more and more data is becoming easily available, and with vast increases in the power of machine learning, there are an increasing number of situations where algorithms—*classifiers*—are used to help decision makers in challenging situations. Examples range from algorithms assisting drivers in cars, to algorithmic methods for determining credit scores, to algorithms helping judges to make sentencing and pretrial decisions in criminal justice. While such *computer-aided decision making* has presented unparalleled levels of accuracy and is becoming increasingly ubiquitous, one of the primary concerns with its widespread adoption is the possibility for such algorithmic methods to lead to structural biases and discriminatory practices (Podesta et al. 2014). A malicious algorithm designer, for instance, might explicitly encode discriminatory rules into a classifier. Perhaps even more problematically, a machine learning method may overfit the data and infer a bias, may inherit a bias from poorly collected data, or may simply be designed to optimize some loss function that leads to discriminatory outcomes.

A well-known instance where this concern has come to light is the debate surrounding the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) tool for recidivism analysis, a classification algorithm that is becoming increasingly widely used in the criminal justice system. Given a series of answers to questions concerning criminal defendants' backgrounds and characteristics, this tool outputs scores from 1 (low risk) to 10 (high risk) estimating how likely they are to recidivate (commit a future crime) or to recidivate violently. According to a recent study by ProPublica (Angwin et al. 2016b), COMPAS and similar risk assessment algorithms are becoming increasingly widely-used throughout the United States; their results are already being shown to judges in nine states during criminal sentencing, and are used in courts nationwide for pretrial decisions such as assigning bail. The ProPublica study, however, found an alarming trend in a set of data collected (Angwin et al. 2016a) concerning individuals' COMPAS results and their actual rates of recidivism over the next two years; in particular, it was found that the scores output by COMPAS lead to a *disparate treatment* of minorities. For instance, in the data collected, African-American defendants *who did not recidivate* were found to be almost twice as likely as white defendants (44.85%, compared to 23.45%) to have been assigned a high risk score (5-10).

Fairness, or non-discrimination, in classification has been studied and debated extensively in the recent past (see (Barocas and Selbst 2016) for an extremely thorough overview); research concerning definitions of fairness in classification dates back to (Pearl 2001) and (Dwork et al. 2012), with more recent definitions tailored to deal with the above-mentioned problems appearing in (Angwin et al. 2016b; Hardt, Price, and Srebro 2016; Chouldechova 2017; Kleinberg, Mullainathan, and Raghavan 2017). To make this setting more concrete, consider some distribution $\mathcal{D}$ from which **individuals** σ are sampled, and consider some **classifier** $\mathcal{C}(\cdot)$ that given some observable features $O(\sigma)$ produces some **outcome**, which later will be used by a decision-maker (DM). The DM is ultimately only interested in the actual **class** $f(\sigma) \in \Psi$ of the individual σ, and their goal is to output some decision $x \in \Omega_{\text{DM}}$ correlated with this actual class. For instance, in the setting of the COMPAS data collected in (Angwin et al. 2016b; 2016a), $\mathcal{D}$ is the distribution over defendants σ, the class $f(\sigma)$ is a bit indicating whether the defendant σ actually commits a crime in the next two years, and the job of the classifier is to output a risk score, which will then be seen and acted upon by a judge. Note that we may without loss of generality assume that the class of the individual is fully determined by σ—situations where the class is probabilistically decided (e.g., at the time of classification, it has yet to be determined whether an individual will or won't recidivate) can be captured by simply including these future coin-tosses needed to determine it into σ, and simply making sure they are not part of the observable features $O(\sigma)$.

Additionally, an individual σ is part of some **group** $g(\sigma) \in \mathcal{G}$—for instance, in the COMPAS setting, the group is the race of the individual. We will refer to the tuple $\mathcal{P} = (\mathcal{D}, f, g, O)$ as a **classification context**. Given such a classification context $\mathcal{P}$, we let $\Psi_{\mathcal{P}}$ denote the range of f, and $\mathcal{G}_{\mathcal{P}}$ denote the range of g. Whenever the classification context $\mathcal{P}$ is clear from context, we drop the subscript; additionally, whenever the distribution $\mathcal{D}, g$ are clear from context, we use $\boldsymbol{\sigma}$ to denote a random variable that is distributed according to $\mathcal{D}$, and $\boldsymbol{\sigma}_X$ to denote the random variable distributed according to $\mathcal{D}$ conditioned on $g(\sigma) = X$.

In this work, we will explore a tension between *fairness for the classifier* and *fairness for the DM*. Roughly speaking, our main result shows that except in "trivial" classification contexts, either the classifier needs to be discriminatory, or a rational decision-maker using the output of the classifier is forced to be discriminatory. Let us turn to describing these two different perspectives on fairness.

Fairness for the Classifier: Fair Treatment The notion of *statistical parity* (Dwork et al. 2012) (which is essentially identical to the notion of causal effect (Pearl 2001)) captures non-discrimination between groups by simply requiring that the output of the classifier be independent (or almost independent) of the group of the individual; that is, for any two groups X and Y, the distributions $\{\mathcal{C}(O(\boldsymbol{\sigma}_X))\}$ and $\{\mathcal{C}(O(\boldsymbol{\sigma}_Y))\}$ are ϵ-close in statistical distance. This is a very strong notion of fairness, and in the above-mentioned context it may not make sense. In particular, if the *base rates* (i.e. the probabilities that individuals from different groups are part of a certain class) are different, we should perhaps not expect the output distribution of the classifier to be the same across groups. Indeed, as the ProPublica article points out, in the COMPAS example, the overall recidivism probability among African-American defendants was 56%, whereas it was 42% among white defendants. Thus, in such situations, one would reasonably expect a classifier to *on average* output a higher risk score for African-American defendants, which would violate statistical parity. Indeed, the issue raised by ProPublica authors was that, even after taking this base difference into account (more precisely, even after conditioning on individuals that did not recidivate), there was a significant difference in how the classifier treated the two races.

The notion of *equalized odds* in (Hardt, Price, and Srebro 2016) formalizes the desiderata articulated by the authors of the ProPublica study (for the case of recidivism) in a general setting by requiring the output of the classifier to be independent of the group of the individuals, *after conditioning on the class of the individuals*.[1] We here consider an approximate version of this notion—which we refer to as **ϵ-fair treatment**—which requires that, for any two groups X and Y and any class c, the distributions

- $\{\mathcal{C}(O(\boldsymbol{\sigma}_X)) \mid f(\boldsymbol{\sigma}_X) = c\}$
- $\{\mathcal{C}(O(\boldsymbol{\sigma}_Y)) \mid f(\boldsymbol{\sigma}_Y) = c\}$

are ϵ-close with respect to some appropriate distance metric to be defined shortly. That is, in the COMPAS example, if we restrict to individuals that actually do not recidivate (respectively, those that do), the output of the classifier ought to be essentially independent of the group of the individual (just as intuitively desired by the authors of the ProPublica study, and as explictly put forward in (Hardt, Price, and Srebro 2016)).

We will use the notion of *max-divergence* to determine the "distance" between distributions; this notion, often found in areas such as differential privacy (see (Dwork 2006)), represents this distance as (the logarithm of) the *maximum multiplicative gap* between the probabilities of some element in the respective distributions. We argue that using such a multiplicative distance is important to ensure fairness between groups or outcomes that may be under-represented in the data.[2] Furthermore, as can be seen in (Morgan and Pass 2018), such a notion is closed under "post-processing": if a classifier $\mathcal{C}$ satisfies ϵ-fair treatment with respect to a context $\mathcal{P} = (\mathcal{D}, f, g, O)$, then for any (possibly probabilistic) function $\mathcal{M}$, $\mathcal{C}'(\cdot) = \mathcal{M}(\mathcal{C}(\cdot))$ will also satisfy ϵ-fair treatment with respect to $\mathcal{P}$. Closure under post-processing is important as we ultimately want the decision-maker to act on the output of the classifier, and we would like the decision-maker's output to be fair whenever they act only on the classifier's output.[3]

As shown in the ProPublica study, the COMPAS classifier does not satisfy ϵ-fair treatment even for somewhat large ϵ. However, several recent works have presented methods to "sanitize" unfair classifiers into ones satisfying ϵ-fair treatment with only a relatively small loss in accuracy (Hardt, Price, and Srebro 2016; Zafar et al. 2017; Morgan and Pass 2018).

Fairness for the Decision-Maker: Rational Fairness. So, classifiers satisfying ϵ-fair treatment with accuracy closely matching the optimal "unfair" classifiers are possible (in fact, classifiers such as COMPAS can be sanitized to satisfy ϵ-fair treatment, without losing too much in accuracy). Additionally, as we have noted, the notion of fair treatment is closed under post-processing, so any mechanism that is applied to the output of the classifier will preserve fair treatment. Thus, intuitively, we would hope that the entire "computer-aided decision making process", where the

[1] Very similar notions of fairness appear also in (Chouldechova 2017; Kleinberg, Mullainathan, and Raghavan 2017) using different names.

[2] For instance, a blatantly discriminatory classifier that positively classified 1% of a group at random, but *only* that group, would have a fair treatment error of 0.01 if we used standard statistical distance, but an infinite max-divergence error.

[3] An earlier approximate definition was proposed in (Kleinberg, Mullainathan, and Raghavan 2017), which simply required that the expectations of the distributions are close; while this is equivalent to our definition for the case of binary outcomes, it is weaker for non-binary outcomes (as in the case of the COMPAS classifier), and this notion is not closed under post-processing.

decision-maker makes use of the classifier's output to make a decision, results in a fair outcome as long as the classifier satisfies fair treatment. Indeed, if the decision-maker simply observes the outcome of the classifier and bases their decision entirely on this outcome, this will be the case (by closure under post-processing).

But the decision-maker is not a machine; rather we ought to think of the DM as a *rational agent*, whose goal is to make decisions that maximize some internal utility function. (For instance, in the context of COMPAS, the DM might be a judge that wants to make sure that defendants that are likely to recidivate are sent to jail, and those who do not are released). As far as we are aware, such a *computer-aided "rational" decision-making* scenario has not yet been studied.

More precisely, we consider a decision-theoretic scenario where individuals σ are sampled from $\mathcal{D}$, the decision-maker gets to see the group $g(\sigma)$ of the individual and the outcome $c = \mathcal{C}(\sigma)$ of the classifier (e.g., the individual's race and risk score), selects some action $x \in \Omega_{\text{DM}}$ (e.g., what sentence to render), and finally receives some utility $u(f(\sigma), x)$ that is only a function of the actual class $f(\sigma)$ (e.g., whether the individual would have recidivated) and their decision x.

Given a classification context $\mathcal{P}$, a classifier $\mathcal{C}$, action space Ω_{DM} and a utility function u, let $\Gamma^{\mathcal{P},\mathcal{C},\Omega_{\text{DM}},u}$ denote the decision problem (i.e., the single-player Bayesian game) induced by the above process. We argue that in a computer-aided decision-making scenario, a natural fairness desideratum for a classifier $\mathcal{C}$ for a context $\mathcal{P}$ is that it should "enable fair rational decision-making". More precisely, we say that a strategy $s : \mathcal{G}_{\mathcal{P}} \times \{0,1\}^* \to \Omega_{\text{DM}}$ for the DM (which chooses an outcome based on the group of the individual and the output of the classifier) is *fair* if the DM ignores the individual's group g and only bases its decision on the output of the classifier—that is, there exists some $s' : \{0,1\}^* \to \Omega_{\text{DM}}$ such that $s(g, o) = s'(o)$. We next say that $\mathcal{C}$ **enables ϵ-approximately fair decision making** (or simply satisfies **ϵ-rational fairness**) with respect to the context $\mathcal{P} = (\mathcal{D}, f, g, O)$ if, for every finite action space Ω_{DM} and every utility function $u : \Psi \times \Omega_{\text{DM}} \to [0,1]$ (i.e., depending on the individual's class and the action selected by the DM), there exists an ϵ-optimal *and fair* strategy s (i.e., a strategy s such that the DM cannot gain more than ϵ in utility by deviating from it) in the induced game $\Gamma^{\mathcal{P},\mathcal{C},\Omega_{\text{DM}},u}$.

Note that if there exists Ω_{DM}, u for which there does not exist some fair ϵ-optimal strategy in the induced game, then there exist situations in which a DM can gain more than ϵ in utility by discriminating between groups, and thus in such situations a rational DM (that cares about "significant" $> \epsilon$ changes in utility) would be *forced* to do so.

Our Main Theorem

Our main result shows that the above-mentioned notions of fairness—which both seem intuitively desirable—are largely incompatible, except in "trivial" cases. In fact, we provide a tight characterization of classification contexts that admit classifiers satisfying ϵ-fair treatment and ϵ-rational fairness.

In these "trivial" cases—for instance, when the features already enable perfect classification, or when the base rates of classes are equal between groups—constructing a fair classifier is possible (and, indeed, usually trivially so) without any significant tradeoffs. However, in non-trivial cases, when these base rates might vary significantly, we show that enforcing fairness will inevitably produce a "predictive disparity" between groups, in that the ability of the outcome of a classifier to predict an individual's true class will need to be sacrificed more in some groups than in others. And, intuitively, this predictive disparity is precisely what causes rational fairness to fail; we show constructively that there are cases where a rational DM will be incentivized to make a more "risky" decision given a group with better predictivity and a "safer" decision given a group with worse predictivity.

The case of binary classes (warm-up). As a warm-up, and to better compare our result to earlier literature, let us start by explaining our characterization for the case of binary classes. We refer to a a classification context $\mathcal{P} = (\mathcal{D}, f, g, O)$ as binary if $\Psi_{\mathcal{P}} = \{0, 1\}$.

We say that a binary classification context $\mathcal{P} = (\mathcal{D}, f, g, O)$ is **ϵ-trivial** if *either* (a) for every class $c \in \{0, 1\}$, the "base rates" of c are ϵ-close with respect to any pair of groups, *or* (b) the observable features enable *perfectly distinguishing* between the two classes. Formally, *either* of the following conditions hold:

- *("almost equal base rates"):* for any two groups X, Y in $\mathcal{G}_{\mathcal{P}}$, and any class $c \in \Psi_{\mathcal{P}}$, the multiplicative distance between $\Pr[f(\boldsymbol{\sigma}_X) = c]$ and $\Pr[f(\boldsymbol{\sigma}_Y) = c]$ is at most ϵ;
- *("perfect distinguishability"):* the distributions $\{O(\boldsymbol{\sigma}) \mid f(\boldsymbol{\sigma}) = 0\}$ and $\{O(\boldsymbol{\sigma}) \mid f(\boldsymbol{\sigma}) = 1\}$ have disjoint support.

Note that if base rates are ϵ-close, there is a trivial classifier that satisfies 0-fair treatment and ϵ-rational fairness: namely, ignore the input and simply output some canonical value. Additionally, note that if the observable features fully determine the class of the individual, there also exists a classifier trivially satisfying 0-fair treatment and 0-rational fairness: simply output the correct class of the individual based on the observable features (which fully determine it by assumption). So ϵ-trivial binary classification contexts admit classifiers satisfying ϵ-fair treatment and ϵ-rational fairness. Our characterization result shows that the above contexts are the only ones which admit them.

Theorem 1. *(Characterizing binary contexts.)* Consider a binary classification context $\mathcal{P} = (\mathcal{D}, f, g, O)$, and let $\epsilon \leq 3/2$ be a constant. Then:

- If $\mathcal{P}$ is ϵ-trivial, there exists a classifier $\mathcal{C}$ satisfying 0-fair treatment and 2ϵ-rational fairness with respect to $\mathcal{P}$.
- If there exists a classifier $\mathcal{C}$ satisfying ϵ-fair treatment and $\epsilon/5$-rational fairness with respect to $\mathcal{P}$, then $\mathcal{P}$ is 4ϵ-trivial.

We note that a similar notion of triviality was considered in (Kleinberg, Mullainathan, and Raghavan 2017; Chouldechova 2017) to obtain related characterizations for binary classification tasks, albeit for different definitions of fairness and "accuracy".

The general case. To deal with the general (i.e., non-binary) case, we need to consider a more general notion of a trivial context. The definition of triviality is actually somewhat different from the definition given for the binary case, but its not hard to see that for this special case the definitions are equivalent.

We say that a classification context $\mathcal{P} = (\mathcal{D}, f, g, O)$ is ϵ**-trivial** if there exists a partition of the set $\Psi_{\mathcal{P}}$ into subsets $\Psi_1, \Psi_2, \dots, \Psi_m$ of classes such that *both* of the following conditions hold:

- *("base-rates conditioned on Ψ_i are close"):* for any $i \in [m]$, for any two groups X, Y in $\mathcal{G}_{\mathcal{P}}$, and any class $c \in \Psi_i$, the multiplicative distance between $\Pr[f(\boldsymbol{\sigma}_X) = c \mid f(\boldsymbol{\sigma}_X) \in \Psi_i]$ and $\Pr[f(\boldsymbol{\sigma}_Y) = c \mid f(\boldsymbol{\sigma}_Y) \in \Psi_i]$ is at most ϵ;
- *("perfect distinguishability between Ψ_i and Ψ_j"):* for any $i \neq j \in [m]$ the distributions $\{O(\boldsymbol{\sigma}) \mid f(\boldsymbol{\sigma}) \in \Psi_i\}$ and $\{O(\boldsymbol{\sigma}) \mid f(\boldsymbol{\sigma}) \in \Psi_j\}$ have disjoint support.

Note that in contrast to the definition given for binary context, the general definition requires that *both* of the above conditions hold (as opposed to just one of them). Note, however, that in case we only have 2 classes, there are only 2 possible partitions of $\Psi_{\mathcal{P}}$: either we have the trivial partition $\Psi_1 = \{0, 1\}$ in which case condition 1 is equivalent to requiring equal base rates, and condition 2 trivially holds; or $\Psi_1 = \{0\}, \Psi_2 = \{1\}$, in which case condition 1 trivially holds, and condition 2 is equivalent to prefect distinguishability between class 0 and class 1.

Once again, if a classification context is ϵ-trivial, there exists a simple classifier that satisfies ϵ-fair treatment and $O(\epsilon)$-rational fairness: given some observable features o, determine which *subgroup* Ψ_i the individual belongs to (which we know can be done by the second requirement), and finally output i. Roughly speaking, this classifier satisfies 0-fair treatment since for any i and any class $c \in \Psi_i$, all individuals in Ψ_i receive the same outcome (namely, i). Rational fairness is a bit more tricky to prove, but roughly speaking follows from the fact that, conditioned on any classification outcome i, the group g of the individual carries "$O(\epsilon)$ information" about the actual class of the individual, and so, by ignoring it, the DM loses little in utility. Our main theorem shows that ϵ-triviality is also a necessary condition:

Theorem 2. *(Full characterization.)* Consider some classification context $\mathcal{P} = (\mathcal{D}, f, g, O)$, let $\epsilon \leq 3/2$ be a constant and let $k = |\Psi_{\mathcal{P}}|$ (i.e., the number of classes). Then:

- If $\mathcal{P}$ is ϵ-trivial, there exists a classifier $\mathcal{C}$ satisfying 0-fair treatment and 2ϵ-rational fairness with respect to $\mathcal{P}$.
- If there exists a classifier $\mathcal{C}$ satisfying ϵ-fair treatment and $\epsilon/5$-rational fairness with respect to $\mathcal{P}$, then $\mathcal{P}$ is $4(k-1)\epsilon$-trivial.

Related Work

Several recent works also show obstacles to achieving fair classifications. Notably, the elegant result of (Kleinberg, Mullainathan, and Raghavan 2017) shows that (in our terminology), for non-trivial *binary* classification problems, there are no classifiers that satisfy ϵ-fair treatment (in fact, an expectation-based relaxation of the notion we consider) as well as a notion of *ϵ-group calibration*—roughly speaking, ϵ-group calibration requires that conditioned on any outcome and group, the distributions of individuals' actual classes are (approximately) "calibrated" according to the outcome. [4] Calibration, however, is best thought of as an "accuracy" notion for the classifier (rather than a fairness notion), and may not always be easy to achieve even without any concern for fairness. (Additionally, the results from (Kleinberg, Mullainathan, and Raghavan 2017) show a weaker bound than those we present here, namely that both of the ϵ-approximate notions they consider in conjunction imply $O(\sqrt{\epsilon})$ difference in base rates or $O(\sqrt{\epsilon})$ prediction error; we present a stronger, asymptotically tight, bound implying either $O(\epsilon)$ difference in base rates or *exact* perfect prediction. However, we note that this is largely due to the fact that the actual definitions employed are incomparable.)

(Chouldechova 2017) presents a similar impossibility result, focusing on binary classification with a binary output. She points out a simple identity (a direct consequence of the definition of conditional probabilities) which implies that, in non-trivial binary classification contexts, and for binary classifiers (i.e., classifiers only outputting a single bit), 0-fair treatment is incompatible with a notion of *perfect* "predictive parity"—namely, that conditioned on the classifier outputting b, the probability that the class is b is independent of the group. While her result only applies in a quite limited setting (binary context, binary classifiers, and only rules out "perfect" fair treatment combined with "perfect" predictive parity), we will rely on an identity similar to hers in one step of our proof. We will also rely on a generalized version of a notion of predictive parity (which deals with non-binary classes, non-binary outcomes, and non-zero error in predictivity) as an intermediate notion within the proof of our main theorem.

As far as we know, no earlier results have considered the effect of having a rational decision-maker act based on the output of the classifier. However, as pointed out to us by an anonymous reviewer, for the case that $\epsilon = 0$, Blackwell's celebrated "comparison of experiments" theorem (Blackwell 1951)[5] be used to show an equivalence between 0-rational fairness and perfect predictive parity, and as such, a Chouldechova's result combined with Blackwell's theorem rules out non-trivial binary classifications admitting classifiers that satisfy 0-fair prediction and 0-rational fairness. Dealing with the case that $\epsilon > 0$, however, is what interests us here: it should be no surprise that "perfect" fairness is impossible, just like "perfect" differential privacy (Dwork 2006) is impossible for any non-trivial task (whereas ϵ-differential privacy where $\epsilon > 0$ is highly possible for many functions of interest!) We highlight that as far as we are

[4] In the COMPAS example, calibration might require that, e.g., of people in each group assigned a risk score of 5, approximately 50% will recidivate, and so forth.

[5] Roughly speaking, this result shows that if a decision-maker can never (i.e., no matter what the utility function is) make use of a signal (in our case, the group of the individual) to improve his utility, then the signal carries no further information than other signals the decision-maker sees (in our case, the output of the classifier).

aware, "approximate" analogues of Blackwell's theorem are not known; in a sense, one of our results—Claim 2—can be viewed as an "approximate" analog of Blackwell's theorem (with a very different type of proof).

Furthermore, to the best of our knowledge, none of the earlier impossibility results consider non-binary classification problems.

Proof Outline

We here provide an outline of the proof of the main theorem. We start by considering just binary classification contexts $\mathcal{P} = (\mathcal{D}, f, g, O)$, and then show how to extend the proof to deal also with non-binary contexts. As mentioned above, for binary contexts, the "if" direction of the theorem (i.e., showing that trivial contexts admit fair classifiers) is immediate. The "only if" direction requires showing that the existence of a classifier $\mathcal{C}$ that satisfies ϵ-fair treatment as well as $\epsilon/5$-rational fairness for a context $\mathcal{P}$ implies that $\mathcal{P}$ is $O(\epsilon)$-trivial. The full proof is deferred to the appendix.

Predictive parity. Towards showing this, we introduce a generalized version of the notion of "predictive parity" considered in (Chouldechova 2017) (which will later also be useful in proving the "if" direction for non-binary classification). Roughly speaking, we say that a classifier satisfies ϵ-predictive parity if, for any two groups X and Y, the following distributions are ϵ-close in multiplicative distance:

- $\{f(\boldsymbol{\sigma}_X) \mid \mathcal{C}(O(\boldsymbol{\sigma}_X)) = c\}$
- $\{f(\boldsymbol{\sigma}_Y) \mid \mathcal{C}(O(\boldsymbol{\sigma}_Y)) = c\}$

That is, the output of the classifier is "equally predictive" of the actual class between groups.

Relating predictive parity and rational fairness. Our first result (which works for all, and not just binary, contexts) shows that rational fairness and predictive parity are intimately connected. First of all, ϵ-predictive parity implies $O(\epsilon)$-rational fairness—intuitively, if a DM could gain by discriminating, then there must exist some output for the classifier for which such a gain is possible, and this contradicts predictive parity. This forward direction turns out to be useful for proving that all ϵ-trivial contexts (even non-binary ones) admit classifiers satisfying ϵ-rational fairness and ϵ-fair treatment; that is, the "if" direction of the theorem (also for non-binary contexts).

More interestingly, we show that $\epsilon/5$ rational fairness (for $\epsilon < 3/2$), *combined with* ϵ'-fair treatment (for any ϵ'), implies ϵ-predictive parity. Intuitively, we show this as follows. Consider some $\mathcal{C}$ that does not satisfy ϵ-predictive parity, yet satisfies $\epsilon/5$-rational fairness and ϵ'-fair treatment. This means there exists some class y^*, groups g, g' and some outcome o such that the prevalence of y^* is significantly higher in group g than in group g' conditioned on the classifier outputting o.

We then construct a very natural game for the DM where every fair strategy has low utility compared to the optimal unfair strategy, which would contradict rational fairness. The action space of the games consists of two actions $\{\mathsf{Risky}, \mathsf{Safe}\}$. If the DM chooses Safe they always receive some fixed utility u^*. On the other hand, if they choose Risky, they receive 1 if the individual's class is y^* and 0 otherwise. That is, playing Risky is good if the individual is "good" (i.e., in class y^*) and otherwise not.

We next show, relying on the fact that $\mathcal{C}$ satisfies fair treatment and the fact that the prevalence of y^* is significantly higher conditioned on the DM getting the signal (o, g) than when getting (o, g'), that, if we set u^* (i.e, the utility of playing Safe) appropriately, the DM can always significantly gain by discriminating between g and g'. The intriguing aspect of this proof is that the optimal "fair" strategy for the DM turns out to be a *mixed* strategy (i.e., a probabilistic strategy) which mixes uniformly between the two actions Risky and Safe.

Simultaneously achieving fair treatment and predictive parity (binary contexts). Given that $O(\epsilon)$-rational fairness combined with (any finite-error) fair treatment implies $O(\epsilon)$-predictive parity, to prove the theorem, it will suffice to show that only trivial contexts admit classifiers that simultaneously satisfy $O(\epsilon)$-fair treatment and $O(\epsilon)$-predictive parity.

Towards showing this, let us first focus on binary classification contexts. We first note that, by the definition of conditional probability, for any $X \in \mathcal{G}_{\mathcal{P}}$, $i, j \in \Psi_{\mathcal{P}}$, and $o \in \Omega_{\mathcal{P}}^{\mathcal{C}}$, the following identity holds:

$$\frac{\Pr[f(\boldsymbol{\sigma}_X) = j \mid \mathcal{C}(O(\boldsymbol{\sigma}_X)) = o]}{\Pr[f(\boldsymbol{\sigma}_X) = i \mid \mathcal{C}(O(\boldsymbol{\sigma}_X)) = o]} \frac{\Pr[\mathcal{C}(O(\boldsymbol{\sigma}_X)) = o \mid f(\boldsymbol{\sigma}_X) = i]}{\Pr[\mathcal{C}(O(\boldsymbol{\sigma}_X)) = o \mid f(\boldsymbol{\sigma}_X) = j]} = \frac{\Pr[f(\boldsymbol{\sigma}_X) = j]}{\Pr[f(\boldsymbol{\sigma}_X) = i]}$$

This identity is basically a generalization of an identity observed in (Chouldechova 2017) for the special case of binary classification tasks and binary classifiers; it relates the conditional probabilities defining fair treatment and predictive parity (the first and second terms on the left, respectively) to the *base rates* of classes between any two groups (the terms on the right).

The same identity as above also holds substituting any $Y \in \mathcal{G}_{\mathcal{P}}$ for X. By applying ϵ-fair treatment and ϵ-predictive parity to these two respective identities, we get that their left-hand sides are 4ϵ-close, and as a consequence we have that the ratios

$$\frac{\Pr[f(\boldsymbol{\sigma}_X) = j]}{\Pr[f(\boldsymbol{\sigma}_X) = i]} \quad \text{and} \quad \frac{\Pr[f(\boldsymbol{\sigma}_Y) = j]}{\Pr[f(\boldsymbol{\sigma}_Y) = i]}$$

are 4ϵ-close. (Note that, to perform these manipulations, it is important that we rely on the multiplicative distance notion.) For the case of binary classification contexts, letting $\alpha_b^g = \Pr[f(\boldsymbol{\sigma}_g) = b]$ denote the "base rate" of class b for group g, this means that the ratios

$$\frac{\alpha_1^X}{\alpha_0^X} = \frac{\alpha_1^X}{1 - \alpha_1^X} \quad \text{and} \quad \frac{\alpha_1^Y}{\alpha_0^Y} = \frac{\alpha_1^Y}{1 - \alpha_1^Y}$$

are 4ϵ-close, and thus we have that the base rates α_1^X, α_1^Y must be 4ϵ-close (and the same for α_0^X, α_0^Y).

But there is a catch. We can only apply the above identity when it is well-defined—that is, when there are no divisions by zero. In other words, we can only apply it if there exists some outcome o such that

$$\Pr[\mathcal{C}(O(\boldsymbol{\sigma})) = o \wedge f(\boldsymbol{\sigma}) = 0] > 0 \quad \text{and}$$
$$\Pr[\mathcal{C}(O(\boldsymbol{\sigma})) = o \wedge f(\boldsymbol{\sigma}) = 1] > 0.$$

If there is no such outcome, $\mathcal{C}$ *perfectly distinguishes* between the two classes, and thus

$$\{O(\boldsymbol{\sigma}) \mid f(\boldsymbol{\sigma}) = 0\} \quad \text{and} \quad \{O(\boldsymbol{\sigma}) \mid f(\boldsymbol{\sigma}) = 1\}$$

must have disjoint support. Hence, in either case, $\mathcal{P}$ is a 4ϵ-trivial context.

Simultaneously achieving fair treatment and predictive parity (general contexts). Dealing with non-binary contexts is quite a bit more involved, and we content ourselves to simply provide a very high-level overview. Consider some $\mathcal{C}$ that satisfies ϵ-fair treatment and ϵ-predictive parity with respect to $\mathcal{P} = (\mathcal{D}, f, g, O)$; our goal is again to show that $\mathcal{P}$ must be $O(\epsilon)$-trivial.

At a high level, we will show either that base rates are ϵ-close or that we can split the set of classes $\Psi_{\mathcal{P}}$ into *proper* subsets Ψ_1, Ψ_2 such that the classifier can perfectly distinguish between these sets of classes. Once we have shown this property, we can next repeatedly rely on it to prove the theorem (more precisely, by recursively splitting up either Ψ_1 or Ψ_2 and applying the same result; formally doing this turns out to be somewhat subtle.)

To prove the above property, our goal is to use the same high-level approach as in the binary case. Assume that there do not exist Ψ_1 and Ψ_2 such that $\mathcal{C}$ can perfectly distinguish between them, and let us show that then the base rates must be close. In order to apply the same argument as in the binary case, we would need to show that for *all* pairs of classes (i, j), the above identity can be applied. If we do this, then we have that, for *all* (i, j), the ratios

$$\frac{\Pr[f(\boldsymbol{\sigma}_X) = j]}{\Pr[f(\boldsymbol{\sigma}_X) = i]} \quad \text{and} \quad \frac{\Pr[f(\boldsymbol{\sigma}_Y) = j]}{\Pr[f(\boldsymbol{\sigma}_Y) = i]}$$

are 4ϵ-close, from which we can conclude that the base rates are 4ϵ-close. However, the fact that $\mathcal{C}$ cannot distinguish between two proper subsets of classes *does not* mean that all classes are "ambiguous" with respect to $\mathcal{C}$ (in the sense that $\mathcal{C}$ cannot perfectly tell them apart, and thus the identity is well-defined). Instead, what we show is that, under the assumption that there do not exist two proper subsets of classes between which $\mathcal{C}$ can perfectly distinguish, we have that, between *any two classes* i and j, there exists a sequence of classes $(i_1, \dots, i_n)$ such that $n \leq k$ (k being the number of classes), $i_1 = i, i_n = j$, and any two *consecutive* classes must be "ambiguous". Ambiguity between classes with respect to $\mathcal{C}$ turns out to be exactly the condition under which the above identity is well defined. At a very high level, we can then perform a "hybrid argument" over the classes in the sequence to still conclude that, for all pairs of classes (i, j), the ratios

$$\frac{\Pr[f(\boldsymbol{\sigma}_X) = j]}{\Pr[f(\boldsymbol{\sigma}_X) = i]} \quad \text{and} \quad \frac{\Pr[f(\boldsymbol{\sigma}_Y) = j]}{\Pr[f(\boldsymbol{\sigma}_Y) = i]}$$

are $4(k-1)\epsilon$-close; this suffices to conclude that the base rates between groups are close.

Acknowledgments

We are extremely grateful to Jon Kleinberg, Silvio Micali and Ilan Komargodski for delightful conversations and insightful suggestions.

References

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016a. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016b. Machine Bias: Risk Assessments in Criminal Sentencing. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Barocas, S., and Selbst, A. 2016. Big Data's Disparate Impact. *California Law Review* 104:671–732.

Blackwell, D. 1951. Comparison of Experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 93–102. University of California Press.

Chouldechova, A. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big data* 5(2):153–163.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, 214–226. New York, NY, USA: ACM.

Dwork, C. 2006. Differential Privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, ICALP'06, 1–12. Berlin, Heidelberg: Springer-Verlag.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, 3323–3331. USA: Curran Associates Inc.

Kleinberg, J. M.; Mullainathan, S.; and Raghavan, M. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, 43:1–43:23.

Morgan, A., and Pass, R. 2018. Achieving Fair Treatment in Algorithmic Classification. In *16th International Conference, TCC 2018, Panaji, India, November 11–14, 2018, Proceedings, Part I*, 597–625.

Pearl, J. 2001. Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, 411–420. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Podesta, J.; Pritzker, P.; Moniz, E. J.; Holdren, J.; and Zients, J. 2014. *Big Data: Seizing Opportunities, Preserving Values*. Executive Office of the President.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness Beyond Disparate Treatment and Disparate Impact: Learning Classification Without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, 1171–1180. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

A Appendix

We present the full proof of our main theorem as an appendix.

B Preliminaries and Definitions

Notation

Conditional probabilities. Given some random variable X and some event E, we let $\Pr[p(X) \mid E]$ denote the probability of a predicate $p(X)$ holding when conditioning the probability space on the event E. If the probability of E is 0, we slightly abuse notation and simply define $\Pr[p(X) \mid E] = 0$.

Multiplicative distance. The following definition of multiplicative distance will be useful to us. We let the **multiplicative distance** $\mu(x, y)$ between two real numbers $x, y \geq 0$ be defined as

$$\mu(x,y) = \begin{cases} \ln\left(\max\left(\frac{x}{y}, \frac{y}{x}\right)\right) & \text{if } x > 0, y > 0 \\ 0 & \text{if } x = y = 0 \\ \infty & \text{otherwise} \end{cases}$$

Classification Contexts

We start by defining classification contexts and classifiers.

Definition 1. A **classification context** $\mathcal{P}$ is denoted by a tuple $(\mathcal{D}, f, g, O)$ such that:

- $\mathcal{D}$ is a probability distribution with some finite support $\Sigma_{\mathcal{P}}$ (the set of all possible **individuals** to classify).
- $f : \Sigma_{\mathcal{P}} \to \Psi_{\mathcal{P}}$ is a surjective function that maps each individual to their **class** in a set $\Psi_{\mathcal{P}}$.
- $g : \Sigma_{\mathcal{P}} \to \mathcal{G}_{\mathcal{P}}$ is a surjective function that maps each individual to their **group** in a set $\mathcal{G}_{\mathcal{P}}$.
- $O : \Sigma_{\mathcal{P}} \to \{0,1\}^*$ is a function that maps each individual to their **observable features**.[6]

We note that f and g are deterministic; this is without loss of generality as we can encode any probabilistic features that f and g may depend on into σ as "unobservable features" of the individual.

Given such a classification context $\mathcal{P}$, we let $\Psi_{\mathcal{P}}$ denote the range of f, and $\mathcal{G}_{\mathcal{P}}$ denote the range of g. Whenever the classification context $\mathcal{P}$ is clear from context, we drop the subscript; additionally, whenever the distribution $\mathcal{D}$ and group function g are clear from context, we use $\boldsymbol{\sigma}$ to denote a random variable that is distributed according to $\mathcal{D}$, and $\boldsymbol{\sigma}_X$ to denote the random variable distributed according to $\mathcal{D}$ conditioned on $g(\sigma) = X$.

A **classifier** $\mathcal{C}$ for a classification context $\mathcal{P} = (\mathcal{D}, f, g, O)$ is simply a (possibly randomized) algorithm. We let $\Omega^{\mathcal{C}}_{\mathcal{P}}$ denote the support of the distribution $\{\mathcal{C}(\boldsymbol{\sigma})\}$.

Fair Treatment

Next we define the notion of *fair treatment*, an approximate version of the notion of "equalized odds" from (Hardt, Price, and Srebro 2016) (which in turn was derived from notions implicit in the ProPublica study (Angwin et al. 2016b)).

Definition 2. We say that a classifier $\mathcal{C}$ satisfies **ϵ-fair treatment** with respect to a context $\mathcal{P} = (\mathcal{D}, f, g, O)$ if, for any groups $X, Y \in \mathcal{G}_{\mathcal{P}}$, any class $c \in \Psi_{\mathcal{P}}$, and any outcome $o \in \Omega^{\mathcal{C}}_{\mathcal{P}}$, we have that

$$\mu(\Pr[\mathcal{C}(O(\boldsymbol{\sigma}_X)) = o \mid f(\boldsymbol{\sigma}_X) = c], \Pr[\mathcal{C}(O(\boldsymbol{\sigma}_Y)) = o \mid f(\boldsymbol{\sigma}_Y) = c]) \leq \epsilon$$

Note that for the case of binary classification tasks and binary classifiers (i.e., when $\Psi_{\mathcal{P}} = \Omega^{\mathcal{C}}_{\mathcal{P}} = \{0,1\}$), fair treatment is equivalent to requiring "similar" false positive and false negative rates.

Rational Fairness

We turn to introducing our notion of "fairness with respect to rational decision-makers". Towards this goal, given a classification context $\mathcal{P}$ and a classifier $\mathcal{C}$, we consider a single-player Bayesian game Γ where individuals σ are sampled from $\mathcal{D}$, the decision-maker (DM) gets to see the group $g(\sigma)$ of the individual and the outcome $o = \mathcal{C}(\sigma)$ of the classifier, and then selects some action $x \in \Omega_{\text{DM}}$. They then receive utility $u(f(\sigma), x)$ that is only a function of the actual class $f(\sigma)$ and their decision x. We let $\Gamma^{\mathcal{P},\mathcal{C},\Omega_{\text{DM}},u}$ denote the Bayesian game induced by the above process (for some action space Ω_{DM} and utility function u). Given such a game $\Gamma^{\mathcal{P},\mathcal{C},\Omega_{\text{DM}},u}$, a **pure strategy for the DM** is a function $s : \mathcal{G}_{\mathcal{P}} \times \Omega^{\mathcal{C}}_{\mathcal{P}} \to \Omega_{\text{DM}}$, and a **mixed strategy** is a probability distribution over pure strategies. In the sequel, we simply use the term "strategy" to refer to mixed strategies.

Definition 3. We say that a strategy s is **ϵ-optimal** in $\Gamma^{\mathcal{P},\mathcal{C},\Omega_{\text{DM}},u}$ where $\mathcal{P} = (\mathcal{D}, f, g, O)$, if for all (g, o) in the support of $\{(g(\boldsymbol{\sigma})), \mathcal{C}(O(\boldsymbol{\sigma}))\}$ and any strategy s', we have:

$$e^{\epsilon}\mathbf{E}[u(f(\boldsymbol{\sigma}), s(g, o)) \mid g(\boldsymbol{\sigma}) = g, \mathcal{C}(\boldsymbol{\sigma}) = o]$$
$$\geq \mathbf{E}[u(f(\boldsymbol{\sigma}), s'(g, o)) \mid g(\boldsymbol{\sigma}) = g, \mathcal{C}(\boldsymbol{\sigma}) = o]$$

That is, a player can never gain more than a factor e^{ϵ} in utility by deviating.$\Gamma^{\mathcal{P},\mathcal{C},\Omega_{\text{DM}},u}$. [7] We turn to defining what it means for a strategy to be fair.

Definition 4. We say that a strategy s for a game $\Gamma^{\mathcal{P},\mathcal{C},\Omega_{\text{DM}},u}$ is **fair** if there exists a function $\tilde{s}$ such that $s(g, o) = \tilde{s}(o)$.

That is, the strategy s does not depend on the group of the individual. As we shall see later on (see Claim 2), the "best" fair strategy s (i.e., a fair strategy that satisfies ϵ-optimality for the smallest ϵ) may need to be a mixed strategy—in fact, we demonstrate a game where there is a significant gap between the best mixed and pure fair strategies. (In our opinion, this is intriguing in its own right, as mixed strategies are typically not helpful in a decision-theoretic—i.e., single-player—setting.)

We finally define what it means for a classifier $\mathcal{C}$ to enable fair decision making.

[6] This is included for generality; for our result, it suffices to take O to be the identity function, as we can show impossibility even for classifiers which may observe every feature of an individual.

[7] Note that we here use the so-called *ex-interim* notion of ϵ-optimality which requires s to be ϵ-close to the optimal strategy even conditioned on the DM having received its type (i.e. (g, o) in our case). This is the most commonly used notion of optimality. We mention that there is also a weaker notion of *ex-ante* ϵ-optimality which only requires s to be optimal *a priori* before seeing the type. A weaker version of our main impossibility result holds also for this notion.

Definition 5. We say that $\mathcal{C}$ **enables ϵ-approximately fair decision making** (or simply satisfies **ϵ-rational fairness**) with respect to the context $\mathcal{P} = (\mathcal{D}, f, g, O)$ if, for every finite action space Ω_{DM} and every utility function $u : \Psi_{\mathcal{P}} \times \Omega_{\text{DM}} \to [0, 1]$, there exists a strategy s that is fair and ϵ-optimal with respect to $\Gamma^{\mathcal{P},\mathcal{C},\Omega_{\text{DM}},u}$.

C Characterizing Fair Classifiers

Our main theorem is a complete characterization of the class of contexts that admit classifiers that simultaneously satisfy fair treatment and rational fairness.

The following notion of "triviality" will characterize contexts admitting such classifiers.

Definition 6. A classification context $\mathcal{P} = (\mathcal{D}, f, g, O)$ is **ϵ-trivial** if there exists a partition of the set $\Psi_{\mathcal{P}}$ into subsets $\Psi_1, \Psi_2, \ldots, \Psi_m$ of classes such that the following conditions hold:

1. For any $i \in [m]$, $c \in \Psi_i$ and any two groups X, Y in $\mathcal{G}_{\mathcal{P}}$, we have that

   $$\mu(\Pr[f(\boldsymbol{\sigma}_X) = c \mid f(\boldsymbol{\sigma}_X) \in \Psi_i], \Pr[f(\boldsymbol{\sigma}_Y) = c \mid f(\boldsymbol{\sigma}_Y) \in \Psi_i]) \leq \epsilon$$

   (i.e., the *base rates* conditioned on Ψ_i are close between groups)
2. For any $i, j \in [m]$ with $i \neq j$, the distributions $\{O(\boldsymbol{\sigma}) \mid f(\boldsymbol{\sigma}) \in \Psi_i\}$ and $\{O(\boldsymbol{\sigma}) \mid f(\boldsymbol{\sigma}) \in \Psi_j\}$ have disjoint support.

Note that if the class space Ψ is binary, triviality means that either the base rates are ϵ-close, or we can perfectly distinguish between the two classes.

Our main characterization theorem shows that a context $\mathcal{P}$ admits classifiers satisfying $O(\epsilon)$-fair treatment and ϵ-rational fairness *if, and only if*, $\mathcal{P}$ is $O(\epsilon)$-trivial.

Theorem 3 (Theorem 2, restated)**.** Consider some classification context $\mathcal{P} = (\mathcal{D}, f, g, O)$ and let $k = |\Psi_{\mathcal{P}}|$ (i.e., the number of classes). Then:

1. For any constant ϵ, if $\mathcal{P}$ is ϵ-trivial, then there exists a classifier $\mathcal{C}$ satisfying 0-fair treatment and 2ϵ-rational fairness with respect to $\mathcal{P}$.
2. For any constant $\epsilon < 3/2$, if there exists a classifier $\mathcal{C}$ satisfying ϵ-fair treatment and $\epsilon/5$-rational fairness with respect to $\mathcal{P}$, then $\mathcal{P}$ is $4(k-1)\epsilon$-trivial.

Note that Theorem 1 from the introduction (i.e., the classification for binary contexts) follows directly as a special case when $k = 2$. Additionally, let us remark that Theorem 3 holds even for a somewhat weaker definition of rational fairness where we only require the existence of a fair ϵ-equilibrium in games with *binary decision spaces* (i.e., $\Omega_{\text{DM}} = \{0, 1\}$), and even if we restrict to this simple and natural subclass of games.

D Proof of Theorem 3

Towards proving Theorem 3, we first define a notion of ϵ-predictive parity and show that a context $\mathcal{P}$ admits classifiers satisfying ϵ-rational fairness and ϵ-fair treatment *if and only if* $\mathcal{P}$ admits a classifier satisfying $O(\epsilon)$-predictive parity and ϵ-fair treatment. (We note that predictive parity is *not* equivalent to rational fairness, but is so for classifiers that also satisfy fair treatment.)

Next, we show that $O(\epsilon)$-triviality characterizes the set of contexts admitting classifiers satisfying ϵ-predictive parity and ϵ-rational fairness. (This second step is interesting in its own right, and can be thought of a significant strengthening of the impossibility result of (Chouldechova 2017), which only showed triviality for the special case when $\epsilon = 0$, the classification context is binary, and the classifier is binary.[8])

Predictive Parity

We first introduce an intermediate notion of approximate "predictive parity" (we are borrowing the name from (Chouldechova 2017), who considered a perfect version of this notion tailored for binary classifiers, where the class is a single bit and the classifier also outputs only a single bit.) Roughly speaking, ϵ-predictive parity requires that the distributions of individuals' *classes*, conditioned on a particular *outcome*, be ϵ-close between groups. We remark that this notion is a strict relaxation of the notion of ϵ-group calibration considered by (Kleinberg, Mullainathan, and Raghavan 2017) (which not only requires that the distribution of the classes be the same between groups conditioned on the outcome o of the classifier, but also that the outcome o "accurately predicts" the class).

Definition 7. We say that a classifier $\mathcal{C}$ satisfies ϵ-**predictive parity** with respect to a context $\mathcal{P} = (\mathcal{D}, f, g, O)$ if, for any groups $X, Y \in \mathcal{G}_{\mathcal{P}}$, any outcome $o \in \Omega^{\mathcal{C}}_{\mathcal{P}}$, and any class $c \in \Psi_{\mathcal{P}}$, we have that

$$\mu(\Pr[f(\boldsymbol{\sigma}_X) = c \mid \mathcal{C}(O(\boldsymbol{\sigma}_X)) = o], \Pr[f(\boldsymbol{\sigma}_Y) = c \mid \mathcal{C}(O(\boldsymbol{\sigma}_Y)) = o]) \leq \epsilon$$

We next show that predictive parity is closely related to rational fairness (at least, when combined with fair treatment). We first show that ϵ-predictive parity implies $O(\epsilon)$-rational fairness.

Claim 1. Let $\mathcal{C}$ be a classifier that satisfies ϵ-predictive parity with respect to a context $\mathcal{P} = (\mathcal{D}, f, g, O)$. Then $\mathcal{C}$ satisfies 2ϵ-rational fairness with respect to $\mathcal{P}$.

Proof. Consider some classifier $\mathcal{C}$ satisfying ϵ-predictive parity with respect to a context $\mathcal{P}$. We will show that $\mathcal{C}$ also satisfies 2ϵ-rational fairness with respect to $\mathcal{P}$.

Let $\mathcal{T}^{\Gamma}_{DM}$ denote the support of $\{(g(\boldsymbol{\sigma})), \mathcal{C}(O(\boldsymbol{\sigma}))\}$ (i.e., the support of the type space in the Bayesian game). Let s^* be an optimal strategy; we may without loss of generality assume that s^* is a pure strategy, since for every type $(g, o) \in \mathcal{T}^{\Gamma}_{DM}$ there exists a deterministic best response. We now show how to modify s^* into a fair strategy s without ever losing too much in expected utility. For every outcome o, pick some g^*_o such that $(g^*_o, o) \in \mathcal{T}^{\Gamma}_{DM}$, and define $s(g, o) = s^*(g^*_o, o)$. Clearly s is fair. We now show that for every pair $(g, o) \in \mathcal{T}^{\Gamma}_{DM}$, the expected utility of playing s^* can never be more than a factor $e^{2\epsilon}$ better than the expected utility of playing s, and thus s is 2ϵ-optimal (as desired).

[8] For this special case, her notion of triviality is a special case of our notion of 0-triviality, which requires that either the base rates are identical for both groups, or one can perfectly predict the class of an individual.

Assume for contradiction that there exists some (g, o) such that $g \neq g_o^*$ and

$$\mathbf{E}[u(f(\boldsymbol{\sigma}), s^*(g, o)) \mid g(\boldsymbol{\sigma}) = g, \mathcal{C}(O(\sigma)) = o]$$

$$> e^{2\epsilon}\mathbf{E}[u(f(\boldsymbol{\sigma}), s(g, o) \mid g(\boldsymbol{\sigma}) = g, \mathcal{C}(O(\boldsymbol{\sigma})) = o]$$

That is,

$$\sum_{y \in \Psi_{\mathcal{P}}} \Pr[f(\boldsymbol{\sigma}_g) = y \mid \mathcal{C}(O(\boldsymbol{\sigma}_g)) = o] u(y, s^*(g, o))$$

$$> e^{2\epsilon} \sum_{y \in \Psi_{\mathcal{P}}} \Pr[f(\boldsymbol{\sigma}_g) = y \mid \mathcal{C}(O(\boldsymbol{\sigma}_g)) = o] u(y, s(g, o))$$

By applying predictive parity (more precisely, that the multiplicative distance between $\Pr[f(\boldsymbol{\sigma}_g) = y \mid \mathcal{C}(O(\boldsymbol{\sigma}_g)) = o]$ and $\Pr[f(\boldsymbol{\sigma}_{g_o^*}) = y \mid \mathcal{C}(O(\boldsymbol{\sigma}_{g_o^*})) = o]$ is at most ϵ) to both the LHS and the RHS (we lose a factor e^ϵ for each application), we get that

$$\sum_{y \in \Psi_{\mathcal{P}}} \Pr[f(\boldsymbol{\sigma}_{g_o^*}) = y \mid \mathcal{C}(O(\boldsymbol{\sigma}_{g_o^*})) = o] u(y, s^*(g, o))$$

$$> \sum_{y \in \Psi_{\mathcal{P}}} \Pr[f(\boldsymbol{\sigma}_{g_o^*}) = y \mid \mathcal{C}(O(\boldsymbol{\sigma}_{g_o^*})) = o] u(y, s(g, o))$$

In other words, (and relying on the fact that $s(g, o) = s^*(g_o^*, o)$),

$$\mathbf{E}[u(f(\boldsymbol{\sigma}), s^*(g, o)) \mid g(\boldsymbol{\sigma}) = g_o^*, \mathcal{C}(O(\sigma)) = o]$$

$$> \mathbf{E}[u(f(\boldsymbol{\sigma}), s^*(g_o^*, o) \mid g(\boldsymbol{\sigma}) = g_o^*, \mathcal{C}(O(\boldsymbol{\sigma})) = o]$$

which is a contradiction since, by assumption, $s^*(g_o^*, o)$ is an optimal move given the type (g_o^*, o). □

As we next show, any classifier that satisfies ϵ-rational fairness *and* ϵ'-fair treatment (for any ϵ') also satisfies $O(\epsilon)$-predictive parity. Intuitively, we show this as follows. Consider some $\mathcal{C}$ that does not satisfy $O(\epsilon)$-predictive parity, yet satisfies ϵ-rational fairness and ϵ'-fair treatment. This means there exists some class y^*, groups g, g' and some outcome o such that the prevalence of y^* is significantly higher in group g than in group g' conditioned on the classifier outputting o.

We then construct a very natural game for the DM where every fair strategy has low utility compared to the optimal unfair strategy. The action space consists of two actions $\{\mathsf{Risky}, \mathsf{Safe}\}$. If the DM chooses Safe they always receive some fixed utility u^*. On the other hand, if they choose Risky, they receive 1 if the individual's class is y^* and 0 otherwise. That is, playing Risky is good if the individual is "good" (i.e., in class y^*) and otherwise not.

Assume there exists some fair strategy s that is ϵ-optimal in this game. We first observe that by ϵ'-fair treatment of $\mathcal{C}$, it must be the case that both (g, o) and (g, o') are in the support of $\{(g(\boldsymbol{\sigma})), \mathcal{C}(O(\boldsymbol{\sigma}))\}$ (i.e., the support of the "type distribution" of the game), and thus optimality of s must hold conditioned on both of them.

We next use the fact that the prevalence of y^* is significantly higher conditioned on the DM getting the signal (o, g) than when getting (o, g'), and thus if we set u^* (i.e, the utility of playing Safe) appropriately, we can ensure that the DM gains by discriminating between g and g' (playing Risky when the group is g, and Safe otherwise). Interestingly, determining by how much a DM can gain by discriminating turns out to be somewhat subtle; it turns out that the "best" fair strategy (i.e., the fair strategy that minimizes the expected utility loss with respect to the optimal strategy) mixes with probability 1/2 between Risky and Safe.

Claim 2. Let $\mathcal{C}$ be a classifier that satisfies $\log\left(\frac{2}{1+e^{-\epsilon/2}}\right)$-rational fairness with respect to a context $\mathcal{P} = (\mathcal{D}, f, g, O)$, as well as ϵ'-fair treatment with respect to $\mathcal{P}$ (for any ϵ'). Then $\mathcal{C}$ satisfies ϵ-predictive parity with respect to $\mathcal{P}$.

Proof. Assume for contradiction that $\mathcal{C}$ satisfies $\log\left(\frac{2}{1+e^{-\epsilon/2}}\right)$-rational fairness and ϵ'-fair treatment (with respect to $\mathcal{P}$), yet does not satisfy ϵ-predictive parity (with respect to $\mathcal{P}$).

Let $\mathcal{T}_{DM}^{\Gamma}$ denote the support of $\{(g(\boldsymbol{\sigma})), \mathcal{C}(O(\boldsymbol{\sigma}))\}$. We first claim that $\mathcal{T}_{DM}^{\Gamma} = \Omega_{\mathcal{P}}^{\mathcal{C}} \times \mathcal{G}_{\mathcal{P}}$. If not, since $\Omega_{\mathcal{P}}^{\mathcal{C}}$ is the support of $\mathcal{C}(O(\boldsymbol{\sigma}))$ (and thus for every $o \in \Omega_{\mathcal{P}}^{\mathcal{C}}$ there is at least one $g \in \mathcal{G}_{\mathcal{P}}$ for which $(o, g) \in \mathcal{T}_{DM}^{\Gamma}$), there must exist an outcome $o \in \Omega_{\mathcal{P}}^{\mathcal{C}}$ and groups $g, g' \in \mathcal{G}_{\mathcal{P}}$ such that $(o, g) \in \mathcal{T}_{DM}^{\Gamma}$ but $(o, g') \notin \mathcal{T}_{DM}^{\Gamma}$. This, however, would mean that there is $y \in \Psi_{\mathcal{P}}$ for which the distributions $\{\mathcal{C}(O(\boldsymbol{\sigma}_g)) \mid f(\boldsymbol{\sigma}_g) = y\}$ and $\{\mathcal{C}(O(\boldsymbol{\sigma}_{g'})) \mid f(\boldsymbol{\sigma}_{g'}) = y\}$ have different supports (and hence infinite max-divergence), as, by definition of $\mathcal{T}_{DM}^{\Gamma}$, o must be in the support of the former for some y but cannot be in the support of the latter for any y. This contradicts ϵ'-fair treatment (for any ϵ') of $\mathcal{C}$.

Next, since $\mathcal{C}$ fails to satisfy ϵ-predictive parity, there exist groups $g, g' \in \mathcal{G}_{\mathcal{P}}$, class $y^* \in \Psi_{\mathcal{P}}$, and an outcome $o \in \Omega_{\mathcal{P}}^{\mathcal{C}}$ such that

$$\frac{\Pr[f(\boldsymbol{\sigma}_g) = y^* \mid \mathcal{C}(O(\boldsymbol{\sigma}_g)) = o]}{\Pr[f(\boldsymbol{\sigma}_{g'}) = y^* \mid \mathcal{C}(O(\boldsymbol{\sigma}_{g'})) = o]} > e^\epsilon$$

Let $\delta > \epsilon$ be such that

$$e^\delta = \frac{\Pr[f(\boldsymbol{\sigma}_g) = y^* \mid \mathcal{C}(O(\boldsymbol{\sigma}_g)) = o]}{\Pr[f(\boldsymbol{\sigma}_{g'}) = y^* \mid \mathcal{C}(O(\boldsymbol{\sigma}_{g'})) = o]}$$

and define the "midpoint" p between these probabilities as

$$p = e^{\delta/2}\Pr[f(\boldsymbol{\sigma}_{g'}) = y^* \mid \mathcal{C}(O(\boldsymbol{\sigma}_{g'})) = o]$$

$$= e^{-\delta/2}\Pr[f(\boldsymbol{\sigma}_g) = y^* \mid \mathcal{C}(O(\boldsymbol{\sigma}_g)) = o]$$

Consider a game where $\Omega_{\text{DM}} = \{\mathsf{Risky}, \mathsf{Safe}\}$, $u(y, \mathsf{Risky}) = 1$ if $y = y^*$ and 0 otherwise and $u(\cdot, \mathsf{Safe}) = p$. The decision-maker's expected utility for choosing Safe is always p; on the other hand:

- Conditioned on (o, g), the decision-maker's expected utility for choosing Risky is

$$\Pr[f(\boldsymbol{\sigma}_g) = y^* \mid \mathcal{C}(O(\boldsymbol{\sigma}_g)) = o] = e^{\delta/2}p$$

- Conditioned on (o, g'), their expected utility for Risky is

$$\Pr[f(\boldsymbol{\sigma}_{g'}) = y^* \mid \mathcal{C}(O(\boldsymbol{\sigma}_{g'})) = o] = e^{-\delta/2}p$$

Consider some fair *pure* strategy s for DM. It must choose either Risky or Safe for both (o, g) and (o, g'). So,

- If s chooses Risky, DM receives $e^{-\delta/2}p$ in expected utility conditioned on (o, g'), whereas they could have received p by instead choosing Safe (thus, they incur a multiplicative loss of $e^{\delta/2}$).
- On the other hand, if s chooses Safe, then the decision-maker receives utility p conditioned on (o, g), whereas they could have received $e^{\delta/2}p$ utility in expectation by instead choosing Risky (again incurring a multiplicative loss of $e^{\delta/2}$).

Thus, we conclude that any fair pure strategy must lose at least a $e^{\delta/2}$ multiplicative factor in utility compared to the optimal (unfair) strategy (which chooses Risky for (o, g) and Safe for (o, g')).

Consider next a fair *mixed* strategy s that chooses Risky with probability p_r (and Safe with probability $1 - p_r$) for both (o, g) and (o, g').

- Conditioned on (o, g'), the decision-maker gets $p_r e^{-\delta/2}p + (1 - p_r)p$ in expected utility, whereas they could have received p by instead choosing Safe. This results in a multiplicative loss of $p_r e^{-\delta/2} + (1 - p_r)$.
- Conditioned on (o, g), the decision-maker gets $p_r e^{\delta/2}p + (1 - p_r)p$ in expected utility, whereas they could have received $e^{\delta/2}p$ utility in expectation by instead choosing Risky, yielding a multiplicative loss of $p_r + (1-p_r)e^{-\delta/2}$.

Thus, when determining the rational strategy that minimizes the loss, we may without loss of generality assume that $p_r \geq 1/2$ (as the case when $p_r \leq 1/2$ is symmetric simply by renaming p_r and $1 - p_r$), and focus on finding the p_r that minimizes

$$p_r e^{-\delta/2} + (1 - p_r) = 1 - p_r(1 - e^{-\delta/2})$$

which happens when p_r is as small as possible, and thus when $p_r = 1/2$. So the optimal mixed rational strategy must be $p_r = 1/2$, and thus has a multiplicative loss of

$$1/2(e^{-\delta/2} + 1) < 1/2(e^{-\epsilon/2} + 1)$$

for both (o, g) and (o, g') compared to the optimal unfair strategy. (In particular, using the expected utilities above, setting $p_r > 1/2$ worsens the multiplicative loss conditioned on (o, g') by decreasing $p_r e^{-\delta/2} + (1 - p_r)$, and setting $p_r < 1/2$ worsens the multiplicative loss conditioned on (o, g) by decreasing $p_r + (1 - p_r)e^{-\delta/2}$.)

Hence, the decision-maker gains at least $\frac{2}{1+e^{-\epsilon/2}}$ utility multiplicatively by switching from any fair mixed strategy to the optimal unfair strategy, and so we conclude the proof with the contradiction that $\mathcal{C}$ cannot satisfy $\log\left(\frac{2}{1+e^{-\epsilon/2}}\right)$-rational fairness with respect to $\mathcal{P}$.

□

We note that, for $\epsilon < 3/2$ (in particular, ϵ less than roughly 1.644), we have that $\log\left(\frac{2}{1+e^{-\epsilon/2}}\right) > \epsilon/5$, which demonstrates the bounds we show in our other results:

Corollary 1. Let $\epsilon \in (0, 3/2)$, and let $\mathcal{C}$ be a classifier that satisfies $\epsilon/5$-rational fairness with respect to a context $\mathcal{P} = (\mathcal{D}, f, g, O)$, as well as ϵ'-fair treatment with respect to $\mathcal{P}$ (for any ϵ'). Then $\mathcal{C}$ satisfies ϵ-predictive parity with respect to $\mathcal{P}$.

An interesting observation (which is not relevant for the sequel of the proof, but nonetheless insightful) which follows from the above proof is that the optimal fair strategy for the DM in the above game is a *mixed* strategy which uniformly mixes between Safe or Risky (each with probability 1/2), whereas any fair pure strategy loses a factor of $e^{\epsilon/2}$ (i.e., significantly more) in utility. (We note, however, that the existence of such a gap between the fair mixed and fair pure strategies can only arise in games where the optimal strategy is unfair: the existence of an optimal fair mixed strategy implies the existence of an optimal fair pure strategy.)

Proof of Theorem 3 (1)

By relying on the fact that predictive parity implies rational fairness, we can now prove the first part of Theorem 3.

Proposition 1 (*Theorem 3 (1).*)**.** If $\mathcal{P} = (\mathcal{D}, f, g, O)$ is an ϵ-trivial context, then there exists a classifier $\mathcal{C}$ that satisfies 0-fair treatment and 2ϵ-rational fairness with respect to $\mathcal{P}$.

Proof. Consider a classifier $\mathcal{C}$ that on input y (in the support of $O(\boldsymbol{\sigma})$) recovers some σ such that $O(\sigma) = y$, and then outputs $f(\sigma)$.

Proving that $\mathcal{C}$ satisfies fair treatment: Consider some $i \in [m]$ and some class $c \in \Psi_i$. We aim to show that for any two groups $X, Y \in \mathcal{G}_\mathcal{P}$ and any outcome o, we have that

$$\Pr[\mathcal{C}(O(\boldsymbol{\sigma}_X)) = o \mid f(\boldsymbol{\sigma}_X) = c]$$
$$= \Pr[\mathcal{C}(O(\boldsymbol{\sigma}_Y)) = o \mid f(\boldsymbol{\sigma}_Y) = c]$$

First note that, by the first condition in the definition of an ϵ-trivial context (i.e., the "equal base rate condition"), it follows that c is in the support of $\{f(\boldsymbol{\sigma}_X)\}$ if and only if it is in the support of $\{f(\boldsymbol{\sigma}_Y)\}$. Next, consider some c in the support of $\{f(\boldsymbol{\sigma}_X)\}$ (and thus also in the support of $\{f(\boldsymbol{\sigma}_Y)\}$). Let i be such that $c \in \Psi_i$. Due to the second condition in the definition of an ϵ-trivial context (i.e., that the distributions $\{O(\boldsymbol{\sigma}) \mid f(\boldsymbol{\sigma}) \in \Psi_j\}$ for $j \in [m]$ have disjoint support), it follows that for every σ such that $f(\sigma) = c$, $\mathcal{C}(O(\sigma))$ always outputs i. Thus,

$$\Pr[\mathcal{C}(O(\boldsymbol{\sigma}_X)) = i \mid f(\boldsymbol{\sigma}_X) = c]$$
$$= \Pr[\mathcal{C}(O(\boldsymbol{\sigma}_Y)) = i \mid f(\boldsymbol{\sigma}_Y) = c] = 1$$

which concludes the proof that $\mathcal{C}$ satisfies 0-fair treatment.

Proving that $\mathcal{C}$ satisfies rational fairness: To show that $\mathcal{C}$ satisfies 2ϵ-rational fairness, we note that, by Claim 1, it suffices to show that $\mathcal{C}$ satisfies ϵ-predictive parity. By the first condition of an ϵ-trivial context, we have that for every $i \in [m]$, $c \in \Psi_i$, and X, Y in $\mathcal{G}_\mathcal{P}$,

$$\mu(\Pr[f(\boldsymbol{\sigma}_X) = c \mid f(\boldsymbol{\sigma}_Y) \in \Psi_i], \Pr[f(\boldsymbol{\sigma}_Y) = c \mid f(\boldsymbol{\sigma}_Y) \in \Psi_i]) \leq \epsilon$$

By the disjoint support assumption, we have that $\mathcal{C}(O(\sigma)) = i$ if and only if $f(\sigma) \in \Psi_i$, thus we have

$$\mu(\Pr[f(\boldsymbol{\sigma}_X) = c \mid \mathcal{C}(O(\boldsymbol{\sigma}_X)) = i], \Pr[f(\boldsymbol{\sigma}_Y) = c \mid \mathcal{C}(O(\boldsymbol{\sigma}_Y)) = i]) \leq \epsilon$$

so $\mathcal{C}$ satisfies ϵ-predictive parity. □

Subgroup Perfect Prediction

To prove the second part of Theorem 3, we introduce some additional notions.

Definition 8. We say that a classifier $\mathcal{C}$ satisfies **subgroup perfect prediction** with respect to context $\mathcal{P}$ if there exists a *proper* subset $\psi \subset \Psi_{\mathcal{P}}$ such that the distributions

$$\{\mathcal{C}(O(\boldsymbol{\sigma})) \mid f(\boldsymbol{\sigma}) \in \psi\} \quad \text{and} \quad \{\mathcal{C}(O(\boldsymbol{\sigma})) \mid f(\boldsymbol{\sigma}) \notin \psi\}$$

have disjoint support.

To characterize classifiers satisfying subgroup perfect prediction, a notion of "ambiguity between classes" will be useful.

Definition 9. Given a classifier $\mathcal{C}$ and context $\mathcal{P}$, we say that classes $i, j \in \Psi_{\mathcal{P}}$ are **ambiguous** (with respect to $\mathcal{C}$ and $\mathcal{P}$) if there exists $o \in \Omega^{\mathcal{C}}_{\mathcal{P}}$ such that $\Pr[f(\boldsymbol{\sigma}) = i \wedge \mathcal{C}(O(\boldsymbol{\sigma})) = o] > 0$ and $\Pr[f(\boldsymbol{\sigma}) = j \wedge \mathcal{C}(O(\boldsymbol{\sigma})) = o] > 0$. We further say that classes $i, j \in \Psi_{\mathcal{P}}$ are n-**ambiguous** if there exists a sequence $(i_0 = i, i_1, i_2, \ldots, i_n = j) \in (\Psi_{\mathcal{P}})^{n+1}$ such that any two consecutive elements i_k and i_{k+1} are ambiguous.

We now have the following useful claim which says that, if a classifier does not satisfy subgroup perfect prediction, then all classes can be connected by a "short" ambiguous sequence.

Claim 3. Consider some classifier $\mathcal{C}$ that does not satisfy subgroup perfect prediction with respect to some context $\mathcal{P} = (\mathcal{D}, f, g, O)$. Then for every pair of classes $i, j \in \Psi_{\mathcal{P}}$, we have that i, j are $m_{i,j}$-ambiguous for some $m_{i,j} \leq |\Psi_{\mathcal{P}}| - 1$.

Proof. Given a context $\mathcal{P}$ and a classifier $\mathcal{C}$, consider a graph G with $n = \Psi_{\mathcal{P}}$ vertices, where we draw an edge between two vertices i, j if i and j are ambiguous. Note that i, j are m-ambiguous if and only if there exists a path of length m connecting them.

We show that the graph must be fully connected if $\mathcal{C}$ does not satisfy subgroup perfect prediction; the proof of the claim immediately follows, as the shortest path between any two nodes in a fully connected graph with n nodes can never be more than $n - 1$.

Assume for the sake of contradiction that G is not fully connected, yet $\mathcal{C}$ does not satisfy subgroup perfect prediction. Then G must have a component ψ disconnected from the remainder of the graph (which can be concretely found by, say, considering the set of all vertices reachable from some class $i \in \Psi_{\mathcal{P}}$). Then we notice that the set of outcomes that can be assigned to individuals with classes in ψ must be entirely disjoint from the set of outcomes that can be assigned to individuals with classes outside ψ; otherwise, there would by definition exist an edge between a vertex in ψ and a vertex in $\Psi_{\mathcal{P}} \setminus \psi$ in the graph, contradicting our assumption that ψ is disconnected from $\Psi_{\mathcal{P}} \setminus \psi$. This contradicts our assumption that $\mathcal{C}$ does not satisfy subgroup perfect prediction. □

The next lemma can be viewed as a weak form of the second part of Theorem 3. (In fact, for the case of binary classification contexts, this lemma on its own directly implies Theorem 1 from the introduction.) Looking forward, we will soon strengthen this lemma by repeatedly applying it to prove the full Theorem 3. In the sequel, we say that a context $\mathcal{P}$ has ϵ-**approximately equal base rates** if for every $X, Y \in \mathcal{G}_{\mathcal{P}}$ and every $i \in \Psi_{\mathcal{P}}$,

$$\mu(\Pr[f(\boldsymbol{\sigma}_X) = i], \Pr[f(\boldsymbol{\sigma}_Y) = i]) \leq \epsilon$$

Lemma 1. Let $\mathcal{P}$ be a context, let $\mathcal{C}$ be a classifier that satisfies ϵ-fair treatment and ϵ-predictive parity with respect to a context $\mathcal{P}$, and let $k = |\Psi_{\mathcal{P}}|$. Then either:

1. $\mathcal{P}$ satisfies $4(k-1)\epsilon$-approximately equal base rates, or
2. $\mathcal{C}$ satisfies subgroup perfect prediction over $\mathcal{P}$.

Proof. Let $\mathcal{C}$ be a classifier that satisfies ϵ-fair treatment and ϵ-predictive parity with respect to $\mathcal{P}$, and let $k = |\Psi_{\mathcal{P}}|$. We will show that either $\mathcal{P}$ satisfies $4(k-1)\epsilon$-approximately equal base rates, or $\mathcal{C}$ satisfies subgroup perfect prediction over $\mathcal{P}$. Towards proving the lemma, let us introduce some additional notation, and prove some helpful propositions:

- Let $\alpha^i_X = \Pr[f(\boldsymbol{\sigma}_X) = i]$ denote the base rate of the class i w.r.t. the group X.
- Let f_i denote the event that $f(\boldsymbol{\sigma}) = i$ and let $\mathcal{C}_o$ denote the event that $\mathcal{C}(O(\boldsymbol{\sigma})) = o$.
- For any $X \in \mathcal{G}_{\mathcal{P}}$, let f^X_i denote the event $f(\boldsymbol{\sigma}_X) = i$ and let $\mathcal{C}^X_o$ denote the event that $\mathcal{C}(O(\boldsymbol{\sigma}_X)) = o$.

The following proposition is a generalization of the identity observed in (Chouldechova 2017).

Proposition 2. Let $i, j \in \Psi_{\mathcal{P}}, o \in \Omega^{\mathcal{C}}_{\mathcal{P}}$, and $i \neq j$. Then, if $\Pr[f^X_i \wedge \mathcal{C}^X_o] > 0$ and $\Pr[f^X_j \wedge \mathcal{C}^X_o] > 0$, we have:

$$\frac{\Pr[\mathcal{C}^X_o \mid f^X_i]}{\Pr[\mathcal{C}^X_o \mid f^X_j]} = \frac{\Pr[f^X_j]}{\Pr[f^X_i]} \frac{\Pr[f^X_i \mid \mathcal{C}^X_o]}{\Pr[f^X_j \mid \mathcal{C}^X_o]}$$

for any $X \in \mathcal{G}_{\mathcal{P}}$.

Proof. First observe that, if $\Pr[f^X_j \wedge \mathcal{C}^X_o] > 0$, then it follows by conditional probability that $\Pr[f^X_j \mid \mathcal{C}^X_o] > 0$, $\Pr[\mathcal{C}^X_o \mid f^X_j] > 0$, and also $\Pr[\mathcal{C}^X_o] > 0$. Then the conclusion follows immediately:

$$\frac{\Pr[\mathcal{C}^X_o \mid f^X_i]}{\Pr[\mathcal{C}^X_o \mid f^X_j]} = \frac{\Pr[\mathcal{C}^X_o \wedge f^X_i]/\Pr[f^X_i]}{\Pr[\mathcal{C}^X_o \wedge f^X_j]/\Pr[f^X_j]}$$

$$= \frac{\Pr[f^X_j]}{\Pr[f^X_i]} \frac{\Pr[f^X_i \mid \mathcal{C}^X_o]\Pr[\mathcal{C}^X_o]}{\Pr[f^X_j \mid \mathcal{C}^X_o]\Pr[\mathcal{C}^X_o]} = \frac{\Pr[f^X_j]}{\Pr[f^X_i]} \frac{\Pr[f^X_i \mid \mathcal{C}^X_o]}{\Pr[f^X_j \mid \mathcal{C}^X_o]}$$

□

We now use the above proposition to get a relationship between the base rate of any two classes that are ambiguous.

Proposition 3. For any two groups $X, Y \in \mathcal{G}_{\mathcal{P}}$, and any two classes $i, j \in \Psi_{\mathcal{P}}$ that are ambiguous w.r.t. $\mathcal{C}$, we have:

$$\mu\left(\frac{\alpha^X_i}{\alpha^Y_i}, \frac{\alpha^X_j}{\alpha^Y_j}\right) \leq 4\epsilon$$

Proof. Consider any two $X, Y \in \mathcal{G}_{\mathcal{P}}$, and any two classes $i, j \in \Psi_{\mathcal{P}}$ that are ambiguous w.r.t. $\mathcal{C}$. By ambiguity, there exists some $o \in \Omega_{\mathcal{P}}^{\mathcal{C}}$ such that

$$\Pr[f_i \wedge \mathcal{C}_o] > 0 \qquad \text{and} \qquad \Pr[f_j \wedge \mathcal{C}_o] > 0.$$

There thus must exist groups g_1, g_2 such that

$$\Pr[f_i^{g_1} \wedge \mathcal{C}_o^{g_1}] > 0 \qquad \text{and} \qquad \Pr[f_j^{g_2} \wedge \mathcal{C}_o^{g_2}] > 0.$$

By fair treatment between the pairs (g_1, X), (g_2, X), (g_1, Y), and (g_2, Y), it follows that

$$\Pr[f_i^X \wedge \mathcal{C}_o^X] > 0, \quad \Pr[f_j^X \wedge \mathcal{C}_o^X] > 0,$$

$$\Pr[f_i^Y \wedge \mathcal{C}_o^Y] > 0, \quad \Pr[f_j^Y \wedge \mathcal{C}_o^Y] > 0.$$

We can thus apply Proposition 2 to conclude:

$$\mu(\Pr[\mathcal{C}_o^X \mid f_i^X]\alpha_i^X \Pr[f_j^X \mid \mathcal{C}_o^X], \Pr[\mathcal{C}_o^X \mid f_j^X]\alpha_j^X \Pr[f_i^X \mid \mathcal{C}_o^X]) = 0$$

By fair treatment, $\mu(\Pr[\mathcal{C}_o^X \mid f_i^X], \Pr[\mathcal{C}_o^Y \mid f_i^Y]) \leq \epsilon$ and $\mu(\Pr[\mathcal{C}_o^X \mid f_j^X], \Pr[\mathcal{C}_o^Y \mid f_j^Y]) \leq \epsilon$, thus[9]

$$\mu(\Pr[\mathcal{C}_o^Y \mid f_i^Y]\alpha_i^X \Pr[f_j^X \mid \mathcal{C}_o^X], \Pr[\mathcal{C}_o^Y \mid f_j^Y]\alpha_j^X \Pr[f_i^X \mid \mathcal{C}_o^X]) \leq 2\epsilon$$

By predictive parity, $\mu(\Pr[f_i^X \mid \mathcal{C}_o^X], \Pr[f_i^Y \mid \mathcal{C}_o^Y]) \leq \epsilon$ and $\mu(\Pr[f_j^X \mid \mathcal{C}_o^X], \Pr[f_j^Y \mid \mathcal{C}_o^Y]) \leq \epsilon$, thus

$$\mu(\Pr[\mathcal{C}_o^Y \mid f_i^Y]\alpha_i^X \Pr[f_j^Y \mid \mathcal{C}_o^Y], \Pr[\mathcal{C}_o^Y \mid f_j^Y]\alpha_j^X \Pr[f_i^Y \mid \mathcal{C}_o^Y]) \leq 4\epsilon$$

But, by Proposition 2 applied to Y (since $\Pr[f_i^Y \wedge \mathcal{C}_o^Y] > 0$ and $\Pr[f_j^Y \wedge \mathcal{C}_o^Y] > 0$), it also follows that:

$$\mu(\Pr[\mathcal{C}_o^Y \mid f_i^Y]\alpha_i^Y \Pr[f_j^Y \mid \mathcal{C}_o^Y], \Pr[\mathcal{C}_o^Y \mid f_j^Y]\alpha_j^Y \Pr[f_i^Y \mid \mathcal{C}_o^Y]) = 0$$

So, dividing the last two expressions[10] (which is possible since $\Pr[f_i^X \wedge \mathcal{C}_o^X] > 0$ and $\Pr[f_j^X \wedge \mathcal{C}_o^X] > 0$) we conclude

$$\mu\left(\frac{\alpha_i^X}{\alpha_i^Y}, \frac{\alpha_j^X}{\alpha_j^Y}\right) \leq 4\epsilon$$

□

Armed with the above proposition, we turn to proving the lemma. Assume for contradiction that $\mathcal{P}$ does not satisfy $4(k-1)\epsilon$-approximately equal base rates, and that $\mathcal{C}$ does not satisfy subgroup perfect prediction over $\mathcal{P}$. Let $n = k - 1$; by Claim 3, we have that, for every pair of classes $i, j \in \Psi_{\mathcal{P}}$, i and j are $m_{i,j}$-ambiguous for some $m_{i,j} \leq n$. By our assumption that $\mathcal{P}$ does not satisfy $4n\epsilon$-equal base rates, there exists some $i \in \Psi_{\mathcal{P}}$ and some $X, Y \in \mathcal{G}_{\mathcal{P}}$ such that

$$\mu(\alpha_i^X, \alpha_i^Y) > e^{4n\epsilon}.$$

Thus at least one of α_i^X and α_i^Y needs to be non-zero, and then by the definition of fair treatment, we have that also the second one must be non-zero. Thus, either

$$\frac{\alpha_i^X}{\alpha_i^Y} > e^{4n\epsilon} \qquad \text{or} \qquad \frac{\alpha_i^X}{\alpha_i^Y} < e^{-4n\epsilon}.$$

We may assume without loss of generality that the former condition holds (as we may otherwise switch X and Y).

By our ambiguity assumptions, for any $j \in \Psi_{\mathcal{P}}$, there is some $m = m_{i,j} \leq n$ and an ambiguous chain $(i_0 = i, i_1, i_2, \ldots, i_m = j) \in (\Psi_{\mathcal{P}})^{m+1}$ so that any two consecutive elements i_s and i_{s+1} are ambiguous; in particular this means that Proposition 3 applies to any consecutive elements in the sequence, and thus for every $s \in [m-1]$

$$\mu\left(\frac{\alpha_{i_s}^X}{\alpha_{i_s}^Y}, \frac{\alpha_{i_{s+1}}^X}{\alpha_{i_{s+1}}^Y}\right) \leq 4\epsilon$$

Hence, iteratively employing Proposition 3, we have

$$\frac{\alpha_{i_1}^X}{\alpha_{i_1}^Y} > e^{4(n-1)\epsilon}, \quad \frac{\alpha_{i_2}^X}{\alpha_{i_2}^Y} > e^{4(n-2)\epsilon}, \quad \ldots,$$

$$\frac{\alpha_{i_m}^X}{\alpha_{i_m}^Y} = \frac{\alpha_j^X}{\alpha_j^Y} > e^{4(n-m)\epsilon} \geq 1$$

Thus, for every $j \in \Psi_{\mathcal{P}}$, we have that

$$\alpha_j^X > \alpha_j^Y$$

which is a contradiction since

$$\sum_{i \in \Psi_{\mathcal{P}}} \alpha_i^X = \sum_{i \in \Psi_{\mathcal{P}}} \alpha_i^Y = 1.$$

□

Proof of Theorem 3 (2)

In this section, we finally prove the second step of Theorem 3.

Proposition 4 (*Theorem 3 (2).*)**.** Let $\epsilon < 3/2$, let $\mathcal{P}$ be a classification context, and let $\mathcal{C}$ be a classifier satisfying ϵ-fair treatment and $\epsilon/5$-rational fairness with respect to $\mathcal{P}$. Then $\mathcal{P}$ is $4(k-1)\epsilon$-trivial, where $k = |\Psi_{\mathcal{P}}|$.

Proof. Consider some classification context $\mathcal{P}$; let $\epsilon < 3/2$ be a constant and let $k = |\Psi_{\mathcal{P}}|$. Assume the existence of a classifier $\mathcal{C}$ satisfying ϵ-fair treatment and $\epsilon/5$-rational fairness with respect to $\mathcal{P}$. We aim to show that $\mathcal{P}$ is $4(k-1)\epsilon$-trivial.

First, note that by Corollary 1, we have that $\mathcal{C}$ also satisfies ϵ-predictive parity. To show that $\mathcal{P}$ is $4(k-1)\epsilon$-trivial, we shall repeatedly apply the following proposition.

Proposition 5. Let $\mathcal{P} = (\mathcal{D}, f, g, O)$ be a context (where $|\Psi_{\mathcal{P}}| = k$) for which there exists a classifier $\mathcal{C}$ satisfying ϵ-fair treatment and ϵ-predictive parity with respect to $\mathcal{P}$.

Let $\Psi_1, \ldots, \Psi_m$ be a partitioning of $\Psi_{\mathcal{P}}$ into subsets such that, for any $i, j \in [m]$ with $i \neq j$, the distributions

$$\{O(\boldsymbol{\sigma}) \mid f(\boldsymbol{\sigma}) \in \Psi_i\} \quad \text{and} \quad \{O(\boldsymbol{\sigma}) \mid f(\boldsymbol{\sigma}) \in \Psi_j\}$$

have disjoint support. Then, either of the following conditions must hold.

- For any two groups X, Y in $\mathcal{G}_{\mathcal{P}}$, and any $i \in [m]$,

  $$\mu(\Pr[f(\boldsymbol{\sigma}_X) = c \mid f(\boldsymbol{\sigma}_X) \in \Psi_i], \Pr[f(\boldsymbol{\sigma}_Y) = c \mid f(\boldsymbol{\sigma}_Y) \in \Psi_i]) \leq 4(k-1)\epsilon$$

  (i.e., X and Y have approximately equal base rates conditioned on each subset of classes Ψ_i).

[9] using the fact that $\mu(ab, c) = x$ and $\mu(b, d) = y$ implies $\mu(ad, c) \leq x + y$

[10] using the fact that $\mu(a/b, c/d) \leq \mu(a, c) + \mu(b, d)$

- There exists some i and some partition of Ψ_i into *proper* subsets Ψ_i^0, Ψ_i^1 such that the distributions $\{O(\boldsymbol{\sigma}) \mid f(\boldsymbol{\sigma}) \in \Psi_i^0\}$ and $\{O(\boldsymbol{\sigma}) \mid f(\boldsymbol{\sigma}) \in \Psi_i^1\}$ have disjoint support.

Proof. Consider $\mathcal{P}, \mathcal{C}, \Psi_1, \ldots, \Psi_m$ satisfying the premise of the proposition. Let $\mathcal{D}_i$ be the distribution over σ obtained by conditioning $\mathcal{D}$ on the event that $f(\sigma) \in \Psi_i$. We claim that $\mathcal{C}$ also satisfies ϵ-fair treatment and ϵ-predictive parity with respect to each context $\mathcal{P}_i = (\mathcal{D}_i, f, g, O)$, as the conditional distributions over which they are defined are unchanged if we restrict to $f(\boldsymbol{\sigma}) \in \Psi_i$. For fair treatment, this is obvious as for any $c \in \Psi_i$, conditioning on $f(\boldsymbol{\sigma}) = c$ is equivalent to conditioning on $f(\boldsymbol{\sigma}) = c \wedge f(\boldsymbol{\sigma}) \in \Psi_i$ (as these events are the same).

For predictive parity, notice that because the distributions $\{O(\boldsymbol{\sigma}) \mid f(\boldsymbol{\sigma}) \in \Psi_i\}$ are mutually disjoint, there is also a partition $\Omega_1, \ldots, \Omega_m$ of the outcome space $\Omega_{\mathcal{P}}^{\mathcal{C}}$ such that $\mathcal{C}(O(\sigma)) \in \Omega_i$ if and only if $f(\sigma) \in \Psi_i$. Thus conditioning on $f(\boldsymbol{\sigma}) \in \Psi_i$ is equivalent to conditioning on $\mathcal{C}(O(\boldsymbol{\sigma})) \in \Omega_i$. We conclude that conditioning on $\mathcal{C}(O(\boldsymbol{\sigma})) = o \wedge f(\boldsymbol{\sigma}) \in \Psi_i$ (whenever this event happens with positive probability) is equivalent to conditioning on $\mathcal{C}(O(\boldsymbol{\sigma})) = o \wedge \mathcal{C}(O(\boldsymbol{\sigma})) \in \Omega_i$ which in turn is equivalent to conditioning on just $\mathcal{C}(O(\boldsymbol{\sigma})) = o$.

Hence, for $\mathcal{C}$ and each context $\mathcal{P}_i$, we can apply Lemma 1, showing that either $\mathcal{P}_i$ has $4(k-1)\epsilon$-approximately equal base rates, or $\mathcal{C}$ satisfies subgroup perfect prediction with respect to $\mathcal{P}_i$. In case all $\mathcal{P}_i$ satisfy $4(k-1)\epsilon$-approximately equal base rates, we are done (we are satisfying condition 1 in the proposition). Otherwise, there must exist some i such that $\mathcal{C}$ satisfies subgroup perfect prediction with respect to $\mathcal{P}_i$; that is, there some proper subset ψ of Ψ_i such that the distributions $\{\mathcal{C}(O(\boldsymbol{\sigma})) \mid f(\boldsymbol{\sigma}) \in \psi\}$ and $\{\mathcal{C}(O(\boldsymbol{\sigma})) \mid f(\boldsymbol{\sigma}) \notin \psi\}$ have disjoint support (when $\boldsymbol{\sigma}$ is defined over $\mathcal{D}_i$), which in turn means that the distributions $\{O(\boldsymbol{\sigma}) f(\boldsymbol{\sigma}) \in \psi\}$ and $\{O(\boldsymbol{\sigma}) \mid f(\boldsymbol{\sigma}) \notin \psi\}$ also have disjoint support. Hence we may partition Ψ_i into $\Psi_i^0 = \Psi_i \setminus \psi$ and $\Psi_i^1 = \psi$ to satisfy the second condition of the proposition (relying on the fact that $\{O(\boldsymbol{\sigma}) \mid f(\boldsymbol{\sigma}) \in \Psi_i\}$ has disjoint support from the support of $\{O(\boldsymbol{\sigma}) \mid f(\boldsymbol{\sigma}) \in \Psi_j\}$ for every $j \neq i$, and thus so will $\{O(\boldsymbol{\sigma}) \mid f(\boldsymbol{\sigma}) \in \Psi_i^b\}$ for $b \in \{0, 1\}$). □

Now, noticing that we may partition $\Psi_{\mathcal{P}}$ into at most $|\Psi_{\mathcal{P}}| = k$ distinct subsets, we can apply Proposition 5 repeatedly at most $k-1$ times (starting with $\Psi_1 = \Psi_{\mathcal{P}}$, every time increasing the number of partitions by one (by replacing Ψ_i with Ψ_i^0 and Ψ_i^1). Thus, when we can no longer further partition some subset Ψ_i, the first condition from the proposition must hold, and thus we have "$4(k-1)\epsilon$-approximately equal base rates conditioned on Ψ_i" for every i. We conclude that $\mathcal{P}$ is $4(k-1)\epsilon$-trivial, which completes the proof of Theorem 3 (2).

□

Concluding the Proof of Theorem 3

Theorem 3 follows as a direct consequence of Proposition 1 and Proposition 4. This concludes the proof of the main theorem.