Tracking and Improving Information in the Service of Fairness

Sumegha Garg* Princeton University
Michael P. Kim† Stanford University
Omer Reingold‡ Stanford University

Abstract

As algorithmic prediction systems have become widespread, fears that these systems may inadvertently discriminate against members of underrepresented populations have grown. With the goal of understanding fundamental principles that underpin the growing number of approaches to mitigating algorithmic discrimination, we investigate the role of *information* in fair prediction. A common strategy for decision-making uses a *predictor* to assign individuals a risk score; then, individuals are selected or rejected on the basis of this score. In this work, we study a formal framework for measuring the *information content* of predictors. Central to the framework is the notion of a *refinement*, first studied by [DF81]. Intuitively, a refinement of a predictor z increases the overall informativeness of the predictions without losing the information already contained in z. We show that increasing information content through refinements improves the downstream selection rules across a wide range of fairness measures (e.g. true positive rates, false positive rates, selection rates). In turn, refinements provide a simple but effective tool for reducing disparity in treatment and impact without sacrificing the utility of the predictions. Our results suggest that in many applications, the perceived "cost of fairness" results from an information disparity across populations, and thus, may be avoided with improved information.

1 Introduction

As algorithmic predictions are increasingly employed as parts of systems that *classify people*, concerns that such classfiers may be *biased* or *discriminatory* have increased correspondingly. These concerns are far from hypothetical; disparate treatment on the basis of *sensitive features*, like race and gender, has been well-documented in diverse algorithmic application domains [BG18, BCZ+16, FK18]. As such, researchers across fields like computer science, machine learning, and economics have responded with many works aiming to address the serious issues of fairness and unfairness that arise in automated decision-making systems.[1]

While most researchers studying algorithmic fairness can agree on the high-level objectives of the field (e.g., *to ensure individuals are not mistreated on the basis of protected attributes; to promote*

*sumeghag@cs.princeton.edu. Part of this work completed while visiting Stanford University.

†mpk@cs.stanford.edu. Part of this work completed while visiting the Weizmann Institute of Science. Supported, in part, by a Google Faculty Research Award, CISPA Center for Information Security, and the Stanford Data Science Initiative.

‡reingold@stanford.edu. Supported by NSF Grant CCF-1763311.

[1] Moritz Hardt's lecture communicates this trend quite succinctly. https://fairmlclass.github.io/1.html#/4

social well-being and justice across populations), there is much debate about how to translate these normative aspirations into a concrete, formal *definition* of what it means for a prediction system to be fair. Indeed, as this nascent field has progressed, the efforts to promote "fair prediction" have grown increasingly divided, rather than coordinated. Exacerbating the problem, [MPB18] identifies that each new approach to fairness makes its own set of assumptions, *often implicitly*, leading to contradictory notions about the right way to approach fairness [Cho17, KMR17]; these inconsistencies add to the "serious challenge for cataloguing and comparing defintions" [MPB18]. Complicating matters further, recent works [LDR+18, CG18] have identified shortcomings of many well-established notions of fairness. At the extreme, these works argue that blindly requiring certain statistical fairness conditions may in fact *harm* the communities they are meant to protect.

The state of the literature makes clear that choosing an appropriate notion of fairness for prediction tasks is a challenging affair. Increasingly, fairness is viewed as a context-dependent notion [SbF+19, HM19], where the "right" notion for a given task should be informed by conversations between computational and social scientists. In the hopes of unifying some of the many directions of research in the area, we take a step back and ask whether there are guiding princples that broadly serve the high-level goals of "fair" prediction, without relying too strongly on any specific notion of fairness. The present work argues that understanding the "informativeness" of predictions needs to be part of any sociotechnical conversation about the fairness of a prediction system. Our main contribution is to provide a technical language with strong theoretical backing to discuss informational issues in the context of fair prediction.

Our contributions. Towards the goal of understanding common themes across algorithmic fairness, we investigate the role of *information* in fair prediction. We study a formal notion of informativeness in predictions and demonstrate that it serves as an effective tool for understanding and improving the utility, fairness, and impact of downstream decisions. In short, we identify that many "failures" of requiring fairness in prediction systems can be explained by an *information disparity* across subpopulations. Further, we provide algorithmic tools that aim to counteract these failures by improving informativeness. Importantly, the framework is not wedded to any specific fairness desideratum and can be applied broadly to prediction settings where discrimination may be a concern. Our main contributions can be summarized as follows:

- First and foremost, we identify informativeness as a key fairness desideratum. We provide information-theoretic and algorithmic tools for reasoning about how much an individual's prediction reveals about their eventual outcome. Our formulation clarifies the intuition that more informative predictions should enable fairer outcomes, across a wide array of interpretations of what it means to be "fair." Notably, *calibration* plays a key technical role in this reasoning; the information-theoretic framework we present relies intimately on the assumption that the underlying predictors are calibrated. Indeed, our results demonstrate a surprising application of these calibration-based methods towards improving parity-based fairness criteria, running counter to the conventional wisdom that calibration and parity are completely at odds with one another.

- In Section 3, we provide a self-contained exposition of the framework we use to study the *information content* of a predictor. Information content formally quantifies the uncertainty over individuals' outcomes given their predictions. Leveraging properties of calibrated predictors, we show that the information content of a predictor is directly related to the *information loss*

between the *true* risk distribution and the *predicted* risk distribution. Therefore, in many cases, information content – a measurable characteristic of the predicted risk distribution – can serve as a proxy for reasoning about the information disparity across groups. To compare the information content of multiple predictors, we need a key concept called a *refinement*; informally, a refinement of a predictor increases the overall information content, without losing any of the original information. Refinements provide the technical tool for reasoning about how to *improve* prediction quality.

- In Section 4, we revisit the question of finding an optimal fair selection rule. For prominent parity-based fairness desiderata, we show that the optimal selection rule can be characterized as the solution to a certain linear program, based on the given predictor. We prove that improving the information content of these predictions via refinements results in a Pareto improvement of the resulting program in terms of utility, disparity, and long-term impact. As one concrete example, if we hold the selection rule's utility constant, then *refining the underlying predictions causes the disparity between groups to decrease*. Additionally, we prove that at times, the *cost* associated with requiring fairness should be blamed on a *lack of information* about important subpopulations, not on the fairness desideratum itself.

- In Section 5, we describe a simple algorithm, `merge`, for incorporating disparate sources of information into a single calibrated predictor. The `merge` operation can be implemented efficiently, both in terms of time and sample complexity. Along the way in our analysis, we introduce the concept of *refinement distance* – a measure of how much two predictors' knowledge "overlaps" – that may be of independent interest.

Finally, a high-level contribution of the present work is to *clarify challenges* in achieving fairness in prediction tasks. Our framework for tracking and improving information is particularly compelling because it does not requires significant technical background in information theory nor algorithms to understand. We hope the framework will facilitate interactions between computational and social scientists to further unify the literature on fair prediction, and ultimately, effect change in the fairness of real-world prediction systems.

1.1 Why information?

We motivate the study of information content in fair prediction by giving an intuitive overview of how information disparities can lead to unfair treatment and impact. We scaffold our discussion around examples from two recent works [LDR+18, CG18] that raise concerns about using broad-strokes statistical tests as the *definition* of fairness. This overview will be informal, prioritizing intuition over technicality; see Section 2 for the formal preliminaries.

We consider a standard prediction setting where a decision maker, who we call the *lender*, has access to a *predictor* $z : \mathcal{X} \to [0, 1]$; from the predicted risk score $z(x)$, the decison maker must choose whether to accept or reject the individual $x \in \mathcal{X}$, i.e. whether to give x a loan or not. For each individual $x \in \mathcal{X}$, we assume there is an associated outcome $y \in \{0, 1\}$ representing if they would default or repay a given loan (0 and 1, respectively). Throughout, we will be focused on *calibrated* predictors. Intuitively, calibration requires that a predicted score of $z(x) = p$ corresponds to the same level of risk, regardless of whether $x \in A$ or $x \in B$. More technically, this means that

we can think of $z(x) = p$ as a conditional *probability*; that is, amongst the individuals who receive score $z(x) = p$, a p-fraction of them end up having $y = 1$. For simplicity, we assume there are two disjoint subpopulations $A, B \subseteq \mathcal{X}$. The works of [LDR$^+$18, CG18] mainly focus on settings where there are material differences between the distribution of y in the populations A and B, arguing that these differences can lead to undesirable outcomes. We argue that even if the true risk of individuals from A and B are identically distributed, differences in the distribution of *predicted* risk scores give rise to the same pitfalls.

A caution against parity. [LDR$^+$18] focuses on notions of fairness that require parity between groups. One notion they study is *demographic parity*, which requires that the selection rate between groups A and B be equal; that is, $\mathbf{Pr}[x \text{ selected} \mid x \in A] = \mathbf{Pr}[x \text{ selected} \mid x \in B]$. Suppose that the majority of applicants come from group A and that on-average, members of A tend to have higher predictions according to z. In such a setting, an unconstrained utility-maximizing lender would give out loans at a higher rate in A than in B. The argument in [LDR$^+$18] against requiring demographic parity goes as follows: a lender who is constrained to satisfy demographic parity must either give out fewer loans in A or more in B; because the lender does not want to give up utility from loaning to A, the constrained lender will give out more loans in B. [LDR$^+$18] argue that in many reasonable settings, the lender will end up loaning to *underqualified* individuals in B who are unlikely to repay; thus, the default rate in B will increase significantly. In their model, this increased default rate translates into *negative impact* on the population B, whose members may go into debt and become even less creditworthy.

A caution against calibration. In general, [CG18] advocates for the use of calibrated score functions paired with threshold selection policies, where an individual is selected if $z(x) > \tau$ for some fixed, group-independent threshold τ. Still, they caution that threshold policies paired with calibrated predictors are not sufficient to guarantee equitable treatment. In particular, suppose that the lender is willing to accept individuals if they have at least 0.7 probability of returning the loan. But now consider a set of calibrated risk scores where the scores are much more confident about A than about B; at the extreme, suppose that for $x \in A$, $z(x) \in \{0, 1\}$ (i.e. perfect predictions) and for $x \in B$, $z(x) = 0.5$ (i.e. uniform predictions). In this case, using a fixed threshold of $\tau = 0.7$ will select every qualified individual in A and none of the individuals from B, even though, by the fact that z is calibrated, half of them were qualified. Worse yet, even if we try to select more members of B, every member of B has a 0.5 probability of defaulting. Indeed, in this example, we cannot distinguish between the individuals in B because they all receive the same score $z(x)$. In other words, we have no *information* within the population B even though the predictor was calibrated.

These examples make clear that when there are actual differences in the risk score distributions between populations A and B, seemingly-natural approaches to ensuring fairness – enforcing parity amongst groups or setting a group-independent threshold – may result in a disservice to the underrepresented population. These works echo a perspective raised by [DHP$^+$12] that emphasizes the distinction between requiring broad-strokes demographic parity as a *constraint* versus stating parity as a *desideratum*. Even if we believe that groups should ideally be treated similarly, defining fairness as satisfying a set of hard constraints may have unintended consequences.

Note that the arguments above relied on differences in the predicted risk scores $z(x)$ for $x \in A$ and $x \in B$, but not the true underlying risk. This observation has two immediate corollaries. On the one hand, in both of these vignettes, if the *predicted* score distributions are different between population

A and B, then such approaches to fairness could still cause harm, *even if the true score distributions are identically distributed*. On the other hand, just because A and B look different according to the predicted scores, they may not actually be different. Intuitively, the difference between the *true* risk distribution and the *observed* risk distribution represents a certain "information loss." Optimistically, if we could somehow improve the informativeness of the predicted scores to reflect the underlying populations more accurately, then the resulting selection rule might exhibit less disparity between A and B in both treatment and impact.

Concretely, suppose we're given a set of predicted risk scores where the scores in A tend to be much more extreme (towards 0 and 1) than those of B. Differences in the risk score distributions such as these can arise for one of two reasons: either individuals from B are *inherently* more stochastic and unpredictable than those in A; or somewhere along the risk estimation pipeline, more information was lost about B than about A. Understanding which story is true can be challenging, if not impossible. Still, in cases where we can reject the hypothesis that certain individuals are inherently less predictable than others, the fundamental question to ask is how to recover the lost information in our predictions. In this work, we provide tools to answer this question.

Refinements. Here, we give a technical highlight of the notion of a *refinement* and the role refinements serve in improving fairness. In Section 3, we introduce the concept of *information content*, $I(z)$ which gives a global measure of how informative a calibrated predictor z is over the population of individuals; intuitively, as $I(z)$ increases, the uncertainty in a typical individual's outcome decreases. The idea that more information in predictions could lead to better utility or better fairness is not particularly surprising. Still, this intuition on its own presents some challenges. For example, suppose we're concerned about minimizing the false positive rate (the fraction of the population where $y = 0$ that were selected). Because I is a *global* measure of uncertainty over all of $\mathcal{X}$, $I(z)$ could be very high due to confidence about a population $S_0 \subseteq \mathcal{X}$ that is very likely to have $y = 0$ ($z(x)$ close to 0), even though z gives very little information ($z(x)$ far from 0 or 1) about the rest of $\mathcal{X}$, which consists of a mix of $y = 0$ and $y = 1$. In this case, a predictor z' with less information ($I(z')$), but better certainty about even a tiny part of the population where $y = 1$ ($z(x)$ close to 1) would enable lower false positive rates (with nontrivial selection rate).

As such, we need another way to reason about what it means for one set of predicted risk scores to have "better information" than the other. Refinements provide the key tool for comparing the information of predictors. Intuitively, a calibrated predictor $\rho : \mathcal{X} \to [0, 1]$ is a refinement of z if ρ hasn't forgotten any of the information contained by z. Formally, we say that ρ *refines* z if $\mathbf{E}_{x\sim\mathcal{X}}[\rho(x) \mid z(x) = v] = v$; this definition is closely related to the idea of calibration in a sense that we make formal in Proposition 3.2.

Refinements allow us to reason about how information influences a broad range of quantities of interest in the context of fair prediction. To give a sense of this, consider the following lemma, which we use to prove our main result in Section 4, but is also independently interesting. The lemma shows that under any fixed selection rate $\beta = \mathbf{Pr}_{x\sim\mathcal{X}}[f(x) = 1]$, the true positive rates, false positive rates, and positive predictive value all improve with a nontrivial refinement.

Lemma. *If ρ is a refinement of z, then for all selection rates $\beta \in [0, 1]$,*

$$\mathrm{TPR}^{\rho}(\beta) \geq \mathrm{TPR}^{z}(\beta), \qquad \mathrm{FPR}^{\rho}(\beta) \leq \mathrm{FPR}^{z}(\beta), \qquad \mathrm{PPV}^{\rho}(\beta) \geq \mathrm{PPV}^{z}(\beta).$$

Intuitively, the lemma shows that by improving information through refinements, mutliple key fairness quantities improve simultaneously. Leveraging this lemma and other properties of refinements and calibration, we show that for many different ways a decision-maker might choose their "optimal" selection rule, the "quality" of the selection rule improves under refinements. We highlight this lemma as one example of the broad applicability of the refinement concept in the context of fair prediction.

Perspective. Disparities in the information content of risk scores may arise for many reasons. The present work clarifies how disparities across groups at early stages of the decision-making pipeline may contribute to disparities in the downstream decisions. In particular, differences in the availability or quality of training data as well as optimization procedures that are tailored for performance in the majority population could contribute to information loss in the minority. The present work highlights the importance of auditing existing risk score predictors for information content across groups, and demonstrates that obtaining informative calibrated predictions can improve fair selection rules, even when the fairness desiderata are based on parity.

Organization. The manuscript is structured as follows: Section 2 establishes notation and covers the necessary preliminaries; Section 3 provides the technical framework for measuring information in predictors; Section 4 demonstrates how improving information content improves the resulting fair selection rules; and Section 5 describes the `merge` algorithm for combining and refining multiple predictors. We conclude with a brief discussion of the context of this work and some directions for future investigation.

1.2 Related Works

The influential work of [DHP+12] provided two observations that are of particular relevance to the present work. First, [DHP+12] emphasized the pitfalls of hoping to achieve "fairness through blindness" by censoring sensitive information during prediction. Second, the work highlighted how enforcing broad-strokes demographic parity conditions – even if desired or expected in fair outcomes – is insufficient to imply fairness. Our results can be viewed as providing further evidence for these perspectives. As discussed earlier, understanding the ways in which fair *treatment* can fail to provide fair *outcomes* [CG18, LDR+18] provided much of the motivation for this work. For a comprehensive overview of the literature on the growing list of approaches to fairness in prediction systems, we recommend the recent encyclopedic survey of [MPB18].

A few recent works have (implicitly or explicitly) touched on the relationship between information and fairness. [CJS18] argues that discrimination may arise in prediction systems due to disparity in predictive power; they advocate for addressing discrimination through data collection. Arguably, much of the work on fairness in online prediction [JKMR16] can be seen as a way to gather information while maintaining fairness. Recently, issues of information and fairness were also studied in unsupervised learning tasks [STM+18]. From the computational economics literature, [KLMR18] presents a simple planning model that draws similar qualitative conclusions to this work, demonstrating the significance of trustworthy information as a key factor in algorithmic fairness.

The idea that better information improves the lender's ability to make useful and fair predictions may seem intuitive under our framing. Interestingly, different framings of prediction and informativeness can lead to qualitatively different conclusions. Specifically, the original work on delayed

impact [LDR+18] suggests that some forms of misestimation (i.e. loss of information) may reduce the potential for harm from applying parity-based fairness notions. In particular, if the lender's predictor z is miscalibrated in a way that underestimates the quality of a group S, then increasing the selection rate beyond the global utility-maximizing threshold may be warranted. In our setting, because we assume that the lender's predictions are calibrated, this type of systematic bias in predictions cannot occur, and more information always improves the resulting selection rule. This discrepancy further demonstrates the importance of group calibration to our notion of information content.

Other works [KRZ19, ILZ19] have investigated the role of *hiding* information through strategic signaling. In such settings, it may be strategic for a group to hide information about individuals in order to increase the overall selection rate for the group. These distinctions highlight the fact that understanding exactly the role of information in "fair" prediction is subtle and also depends on the exact environment of decision-making. We further discuss how to interpret our theorems as well as the importance of faithfully translating fairness desiderata into mathematical constraints/objectives in Section 6.

The present work can also be viewed as further investigating the tradeoffs between calibration and parity. Inspired by investigative reporting on the "biases" of the COMPAS recidivism prediction system [ALMK16], the incompatability of calibration and parity-based notions of fairness has received lots of attention in recent years [Cho17, KMR17, PRW+17]. Perhaps counterintuitively, our work shows how to leverage properties of calibrated predictors to improve the disparity of the eventual decisions. At a technical level, our techniques are similar in flavor to the those of [HKRR18], which investigates how to strengthen group-level calibration as a notion of fairness; we discuss further connections to [HKRR18] in Section 6.

Outside the literature on fair prediction, our notions of information content and refinements are related to other notions from the fields of online forecasting and information theory. In particular, the idea of refinements was first introduced in [DF81]. The concept of information content of calibrated predictions is related to ideas from the forecasting literature [GBR07, GR07], including *sharpness* and *proper scoring rules* [Bri50]. The concept of a refinement of a calibrated predictor can be seen as a special case of Blackwell's informativeness criterion [Bla53, Cré82, DF81].

2 Preliminaries

Basic notation. Let $\mathcal{X}$ denote the domain of individuals and $\mathcal{Y} = \{0, 1\}$ denote the binary outcome space. We assume that individuals and their outcomes are jointly distributed according to $\mathcal{D}$ supported on $\mathcal{X} \times \mathcal{Y}$. Let $x, y \sim \mathcal{D}$ denote an independent random draw from $\mathcal{D}$. For a subpopulation $S \subseteq \mathcal{X}$, we use the shorthand $x, y \sim \mathcal{D}_S$ to be the data distribution conditioned on $x \in S$, and $x \sim S$ to denote a random sample from the marginal distribution over $\mathcal{X}$ conditioned on membership in S.

Predictors. A basic goal in learning is to find a *classifier* $f : \mathcal{X} \to \{0, 1\}$ that given $x \sim \mathcal{X}$ drawn from the marginal distribution over individuals, accurately predicts their outcome y. One common strategy for binary classification first maps individuals to a real-valued *score* using a *predictor* $z : \mathcal{X} \to [0, 1]$ and then selects individuals on the basis of this score. We denote by

$\text{supp}(z)$ the support of z. We denote by $p^* : \mathcal{X} \to [0,1]$ the *Bayes optimal predictor*, where $p^*(x) = \mathbf{Pr}_{\mathcal{D}}\left[y = 1 \;\middle|\; x\right]$ represents the inherent uncertainty in the outcome given the individual. Equivalently, for each individual $x \in \mathcal{X}$, their outcome y is drawn independently from $\text{Ber}(p^*(x))$, the Bernoulli distribution with expectation $p^*(x)$. While we use $[0,1]$ to denote the codomain of predictors, throughout this work, we assume that the set of individuals is finite and hence, the support of any predictor is a discrete, finite subset of the interval.

Risk score distributions. Note that there is a natural bijection between predictors and *score distributions*. A predictor z, paired with the marginal distribution over $\mathcal{X}$, induces a score distribution, which we denote $\mathcal{R}^z$, supported on $[0,1]$, where the probability density function is given as $\mathcal{R}^z(v) = \mathbf{Pr}_{x\sim\mathcal{X}}\left[z(x) = v\right]$. For a subpopulation $S \subseteq \mathcal{X}$, we denote by $\mathcal{R}^z_S$ the score distribution conditioned on $x \in S$.

Calibration. A useful property of predictors is called *calibration*, which implies that the scores can be interpreted meaningfully as the probability that an individual will result in a positive outcome. Calibration has been studied extensively in varied contexts, notably in forecasting and online prediction (e.g. [FV98]), and recently as a fairness desideratum [KMR17,PRW$^+$17,HKRR18, CG18]; the definition we use is adapted from the fairness literature.

Definition (Calibration)**.** *A predictor $z : \mathcal{X} \to [0,1]$ is* calibrated *on a subpopulation $S \subseteq \mathcal{X}$ if for all $v \in \text{supp}(z)$,*

$$\mathop{\mathbf{Pr}}_{x,y\sim\mathcal{D}_S}\left[y = 1 \;\middle|\; z(x) = v\right] = v.$$

For convenience when discussing calibration, we use the notation $S_{z(x)=v} = \{x \in S : z(x) = v\}$. Note that we can equivalently define *calibration with respect to the Bayes optimal predictor*, where z is calibrated on S if for all $v \in \text{supp}(z)$, $\mathbf{E}_{x\sim S_{z(x)=v}}\left[p^*(x)\right] = v$. Operationally in proofs, we end up using this definition of calibration. This formulation also makes clear that p^* is calibrated on every subpopulation.

Parity-based fairness. As a notion of fairness, calibration aims to ensure similarity between predictions and the true outcome distribution. Other fairness desiderata concern disparity in prediction between subpopulations on the basis of a sensitive attribute. For simplicity, we will imagine individuals are partitioned into two subpopulations $A, B \subseteq \mathcal{X}$; we will overload notation and use $\{A, B\}$ to denote the names of the attributes as well. We let $\mathcal{A} : \mathcal{X} \to \{A, B\}$ map individuals to their associated attribute. The most basic notion of parity is *demographic parity* (also sometimes called *statistical parity* in the literature), which states that the selection rate of individuals should be independent of the sensitive attribute.

Definition (Demographic parity [DHP$^+$12])**.** *A selection rule $f : \mathcal{X} \to \{0,1\}$ satisfies* demographic parity *if*

$$\mathop{\mathbf{Pr}}_{x\sim\mathcal{X}}\left[f(x) = 1 \;\middle|\; \mathcal{A}(x) = A\right] = \mathop{\mathbf{Pr}}_{x\sim\mathcal{X}}\left[f(x) = 1 \;\middle|\; \mathcal{A}(x) = B\right].$$

One critique of demographic parity is that the notion does not take into account the actual qualifications of groups (i.e. no dependence on y). Another popular parity-based notion, called *equalized opportunity*, addresses this criticism by enforcing parity of *false negative rates* across groups.

Definition (Equalized opportunity [HPS16])**.** *A selection rule* $f : \mathcal{X} \to \{0, 1\}$ *satisfies* equalized opportunity *if*

$$\mathop{\mathbf{Pr}}_{x,y\sim\mathcal{D}}\left[f(x) = 0 \;\middle|\; y = 1,\ \mathcal{A}(x) = A\right] = \mathop{\mathbf{Pr}}_{x,y\sim\mathcal{D}}\left[f(x) = 0 \;\middle|\; y = 1,\ \mathcal{A}(x) = B\right]$$

In addition to these fairness concepts, the following properties of a selection rule will be useful to track. Specifically, we define the true positive rate (TPR), false positive rate (FPR), and positive predictive value (PPV).

$$\begin{aligned}
\mathrm{TPR}(f) &= \mathop{\mathbf{Pr}}_{x,y\sim\mathcal{D}}\left[f(x) = 1 \;\middle|\; y = 1\right] \\
\mathrm{FPR}(f) &= \mathop{\mathbf{Pr}}_{x,y\sim\mathcal{D}}\left[f(x) = 1 \;\middle|\; y = 0\right] \\
\mathrm{PPV}(f) &= \mathop{\mathbf{Pr}}_{x,y\sim\mathcal{D}}\left[y = 1 \;\middle|\; f(x) = 1\right]
\end{aligned}$$

3 Measuring information in binary prediction

In this section, we give a self-contained exposition of a formal notion of information content in calibrated predictors. These notions have been studied extensively in the forecasting literature (see [GBR07,GR07] and references therein), but are less common in the literature on computational and statistical learning theory. Our notion of information content can be derived from the Brier scoring rule [Bri50].

In the context of binary prediction, a natural way to measure the "informativeness" of a predictor is by the uncertainty in an individual's outcome given their score. We quantify this uncertainty using *variance*.[2] For $p \in [0, 1]$, the variance of a Bernoulli random variable with expected value p is given as $\mathbf{Var}(\mathrm{Ber}(p)) = p \cdot (1-p)$. Note that variance is a strictly concave function in p and is maximized at $p = 1/2$ and minimized $p \in \{0, 1\}$; that is, a Bernoulli trial with $p = 1/2$ is maximally uncertain whereas a trial with $p = 0$ or $p = 1$ is perfectly certain. Consider a random draw $x, y \sim \mathcal{D}$. If z is a calibrated predictor, then given x and $z(x)$, the conditional distribution over y follows a Bernoulli distribution with expectation $z(x)$. This observation suggests the following defintion.

Definition (Information content)**.** *Suppose for* $S \subseteq \mathcal{X}$*,* $z : \mathcal{X} \to [0, 1]$ *is calibrated on* S*. The* information content *of* z *on* S *is given as*

$$I_S(z) = 1 - 4 \cdot \mathop{\mathbf{E}}_{x\sim S}\left[z(x)(1 - z(x))\right].$$

For a calibrated z, we use $I(z) = I_{\mathcal{X}}(z)$ to denote the "information content of z". The factor 4 in the definition of information content acts as a normalization factor such that $I(z) \in [0, 1]$. At the extremes, a perfectly informative predictor has information content 1, whereas a calibrated predictor that always outputs $1/2$ has 0 information.

[2] Alternatively, we could measure uncertainty through Shannon entropy (in fact, any function that admits a Bregman divergence). The generality of the approach is made clear in [GR07]. In Appendix A, we show that notions of information that arise from Shannon entropy are effectively interchangeable with those that arise from variance. We elect to work with variance in the main body primarily because it simplifies the analysis in Section 5.

This formulation of information content as uncertainty in a binary outcome is intuitive in the context of binary classification. In some settings, however, it may be more instructive to reason about risk score distributions directly. A conceptually different approach to measuring informativeness of a risk score distribution might track the uncertainty in the true (Bayes optimal) risk, given the predicted risk score.

Consider a random variable $P^*_{z(x)=v}$ that takes value $p^*(x)$ for x sampled from the individuals with score $z(x) = v$; that is, $P^*_{z(x)=v}$ equals the true risk for an individual sampled amongst those receiving predicted risk score v. Again, we could measure the uncertainty in this random variable by tracking its variance given $z(x) = v$; the higher the variance, the less information the risk score distribution $\mathcal{R}_z$ provides about the true risk score distribution $\mathcal{R}_{p^*}$. Recall, for a predictor z that is calibrated on $S \subseteq \mathcal{X}$ and score $v \in [0, 1]$, we let $S_{z(x)=v} = \{x \in S : z(x) = v\}$. Consider the variance in P^*_v given as

$$\begin{aligned}\mathbf{Var}\left[P^*_{z(x)=v}\right] &= \mathop{\mathbf{Var}}_{x\sim S_{z(x)=v}}\left[p^*(x)\right] \\ &= \mathop{\mathbf{E}}_{x\sim S_{z(x)=v}}\left[\left(p^*(x) - v\right)^2\right].\end{aligned}$$

We define the *information loss* by taking an expectation of this conditional variance over the score distribution induced by z.

Definition (Information loss)**.** *For $S \subseteq \mathcal{X}$, suppose a predictor $z : \mathcal{X} \to [0, 1]$ is calibrated on S. The* information loss *of z on S is given as*

$$L_S(p^*; z) = 4 \cdot \mathop{\mathbf{E}}_{\substack{v\sim\mathcal{R}^z_S \\ x\sim S_{z(x)=v}}}\left[\left(p^*(x) - v\right)^2\right]$$

Again, the factor 4 is simply to normalize the information loss into the range $L(p^*; z) \in [0, 1]$. This loss is maximized when p^* is a 50:50 mix of $\{0, 1\}$ but z always predicts $1/2$; the information loss is minimized for $z = p^*$. We observe that this notion of information loss is actually proportional to the expected squared error of z with respect to p^*; that is,

$$\begin{aligned}\mathop{\mathbf{E}}_{\substack{v\sim\mathcal{R}^z_S \\ x\sim S_{z(x)=v}}}\left[\left(p^*(x) - v\right)^2\right] &= \sum_{v\in\mathrm{supp}(z)} \mathcal{R}^z_S(v) \cdot \mathop{\mathbf{E}}_{x\sim S_{z(x)=v}}\left[\left(p^*(x) - v\right)^2\right] \\ &= \mathop{\mathbf{E}}_{x\sim S}\left[\left(p^*(x) - z(x)\right)^2\right]\end{aligned}$$

Thus, for calibrated predictors, we can interpret the familiar squared loss between a predictor and the Bayes optimal predictor as a notion of information loss.

Connecting information content and information loss. As we defined them, information content and information loss seem like conceptually different ways to measure uncertainty in a predictor. Here, we show that they actually capture the same notion. In particular, we can express information loss of z as the difference in information content of p^* and that of z.

Proposition 3.1. *Let $p^* : \mathcal{X} \to [0, 1]$ denote the Bayes optimal predictor. Suppose for $S \subseteq \mathcal{X}$, $z : \mathcal{X} \to [0, 1]$ is calibrated on S. Then*

$$L_S(p^*; z) = I_S(p^*) - I_S(z).$$

Proof. The proof follows by expanding the information loss and rearranging so that, assuming z is calibrated, terms cancel. As notational shorthand, let $S_v = S_{z(x)=v}$ and let $\bar{p} = \mathbf{E}_{x\sim S}[p^*(x)] = \mathbf{E}_{x\sim S}[z(x)]$. Thus, we can rewrite the information loss $L_S(p^*; z)$ as follows.

$$\begin{aligned}
4 \cdot \mathop{\mathbf{E}}_{x\sim S}\left[(p^*(x) - z(x))^2\right] &= 4 \cdot \mathop{\mathbf{E}}_{x\sim S}\left[p^*(x)^2 + z(x)^2 - 2p^*(x)\cdot z(x)\right] \\
&= 4 \cdot \sum_{v\in\text{supp}(z)} \mathcal{R}_S^z(v) \cdot \left(\mathop{\mathbf{E}}_{x\sim S_v}\left[p^*(x)^2\right] + v^2 - 2\mathop{\mathbf{E}}_{x\sim S_v}\left[p^*(x)\right]\cdot v\right) \\
&= 4 \cdot \sum_{v\in\text{supp}(z)} \mathcal{R}_S^z(v) \cdot \mathop{\mathbf{E}}_{x\sim S_v}\left[p^*(x)^2 - v^2\right] \qquad (1) \\
&= 4 \cdot \mathop{\mathbf{E}}_{x\sim S}\left[p^*(x)^2 - z(x)^2\right] \\
&= \left(1 - 4 \cdot \mathop{\mathbf{E}}_{x\sim S}\left[\bar{p} - p^*(x)^2\right]\right) - \left(1 - 4 \cdot \mathop{\mathbf{E}}_{x\sim S}\left[\bar{p} - z(x)^2\right]\right) \\
&= I_S(p^*) - I_S(z)
\end{aligned}$$

where (1) follows because $\mathbf{E}_{x\sim S_v}[p^*(x)] = v$ under calibration. □

Because the information loss is a nonnegative quantity, Proposition 3.1 also formalizes the inutition that the Bayes optimal predictor is the most informative predictor; for $z \neq p^*$, $I(z) < I(p^*) \leq 1$. Ideally, in order to evaluate the information disparity across groups, we would compare the information lost from p^* to z across A and B. But because the definition of information loss depends on the true score distribution p^*, in general, it's impossible to directly compare the loss. Still, if we believe that $p^*(x)$ is similarly distributed across $x \sim A$ and $x \sim B$, then measuring the information contents $I_A(z)$ and $I_B(z)$ – properties of the *observed* risk scores – allows us to directly compare the loss.

3.1 Incorporating information via refinements

We have motivated the study of informativeness in prediction with the intuition that as information content improves, so too will the resulting fairness and utility of the decisions derived from the predictor. Without further assumptions, however, this line of reasoning turns out to be overly-optimistic. For instance, consider a setting where the expected utility of lending to individuals is positive if $z(x) > \tau$ for some fixed threshold τ (and negative otherwise). In this case, information about individuals whose $p^*(x)$ is significantly below τ is not especially useful. Figure 1 gives an example of two predictors, each calibrated to the same $p^*(x)$, where $I(z') > I(z)$, but z is preferable. At a high-level, the example exploits the fact that information content $I(z)$ is a global property of z, whereas the quantities that affect the utility and fairness directly, like PPV or TPR are *conditional* properties. Even if $I(z') > I(z)$, it could be that z' has lost information about an important subpopulation compared to z, compensating with lots of information about the unqualified individuals.

Still, we would like to characterize ways in which more information is definitively "better." Intuitively, more information is better when we don't have to give up on the information in the current predictor, but rather refine the information contained in the predictions further. The following definition, equivalent to a notion proposed in [DF81], formalizes this idea.

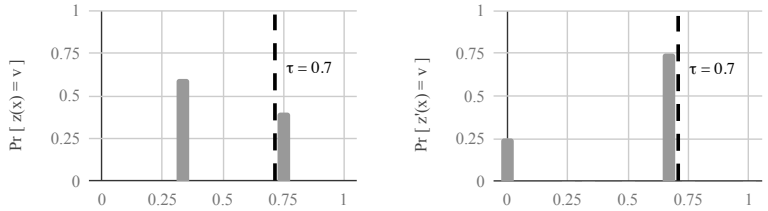

Figure 1: Comparing information content directly is insufficient to compare predictors' utility. *Two predictors* $z, z' : \mathcal{X} \to [0,1]$ *are each calibrated to* $p^* : \mathcal{X} \to \{0,1\}$ *with* $\mathbf{E}[p^*(x)] = 1/2$. *In* z, $\mathbf{Pr}[z(x) = 1/3] = 3/5$ *and* $\mathbf{Pr}[z(x) = 3/4] = 2/5$*; in* z', $\mathbf{Pr}[z(x) = 0] = 1/4$ *and* $\mathbf{Pr}[z(x) = 2/3] = 3/4$. $I(z') > I(z)$*, but* z *achieves better utility than* z' *whenever* $2/3 < \tau < 3/4$.

Definition (Refinement)**.** *For* $S \subseteq \mathcal{X}$*, suppose* $z, z' : \mathcal{X} \to [0,1]$ *are calibrated on* S. z' *is a* refinement *of* z *on* S *if for all* $v \in \text{supp}(z)$*,*

$$\mathop{\mathbf{E}}_{x \sim S_{z(x)=v}} \left[z'(x)\right] = v.$$

That is, we say that z' *refines* z if z' maintains the same expectation over the level sets $S_{z(x)=v}$. To understand why this property makes sense in the context of maintaining information from z to z', suppose the property was violated: that is, there is some $v \in \text{supp}(z)$ such that $\mathbf{E}_{x \sim S_{z(x)=v}}[z'(x)] \neq v$. This disagreement provides evidence that z has some consistency with the true risk that z' is lacking; because z is calibrated, $\mathbf{E}_{x \sim S_{z(x)=v}}[z(x)] = v = \mathbf{E}_{x \sim S_{z(x)=v}}[p^*(x)]$. In other words, even if z' has greater information content, it may not be consistent with the content of z.

Another useful perspective on refinements is through measuring the information on each of the sets $S_{z(x)=v}$. Restricted to $S_{z(x)=v}$, z has minimal information content – its predictions are constant – whereas z' may vary. Because z' is calibrated and maintains the expectation over $S_{z(x)=v}$, we can conclude that $I_{S_{z(x)=v}}(z) \leq I_{S_{z(x)=v}}(z')$ for each of the partitions.

This perspective highlights the importance of requiring *calibration* in the definition of refinements. Indeed, because a refinement is a calibrated predictor, refinements cannot make arbitrary distinctions in predictions, so any additional distinctions on the level sets of the original predictor must represent true variability in p^*. We draw attention to the similarity between the definition of a refinement and the definition of calibration. In particular, if z' is a refinement of z, then z is not only calibrated with respect to p^*, but also with respect to z'; stated differently, p^* is a refinement of every calibrated predictor. Indeed, one way to interpret a refinement is as a "candidate" Bayes optimal predictor. Carrying this intuition through, we note that the only property of p^* we used in the proof of Proposition 3.1 is that it is a refinement of a calibrated z. Thus, we can immediately restate the proposition in terms of generic refinements.

Proposition 3.2. *Suppose for* $S \subseteq \mathcal{X}$*,* $z, z' : \mathcal{X} \to [0,1]$ *are calibrated on* S*. If* z' *refines* z *on* S*, then* $L_S(z'; z) = I_S(z') - I_S(z)$*.*

This characterization further illustrates the notion that a refinement z' could plausibly be the true risk given the information in the current predictions z. In particular, because $L_S(z'; z) > 0$, we get $I_S(z') > I_S(z)$ for any refinement $z' \neq z$.

In the context of fair prediction, we want to ensure that the information content on specific protected subpopulations does not decrease. Indeed, in this case, it may be important to ensure that the

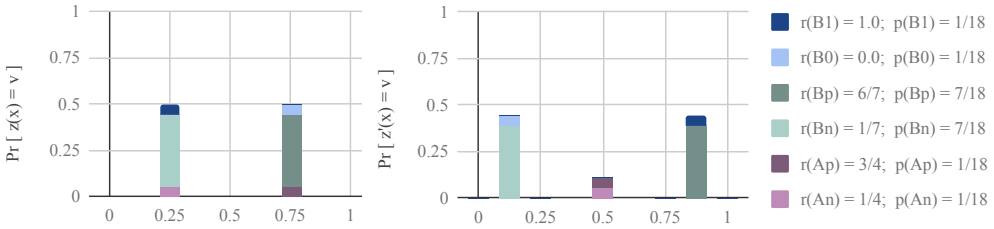

Figure 2: Per-group refinement is necessary to maintain information for each group. *Let $A = A_n \cup A_p$ and $B = B_n \cup B_p \cup B_0 \cup B_1$ where $r(S) = \mathbf{E}_{x \sim S}[p^*(x)]$ and $p(S) = \mathbf{Pr}_{x \sim \mathcal{X}}[x \in S]$. The two predictors $z, z' : \mathcal{X} \to [0,1]$ are each calibrated on A and B. Note that z' refines z overall, but has lost all information about A.*

predictions are refined, not just overall, but also on the sensitive subpopulations. In Figure 2, we illustrate this point by showing two predictors $z, z' : \mathcal{X} \to [0,1]$ that are each calibrated on two subpopulations $A, B \subseteq \mathcal{X}$; z' refines z on $\mathcal{X}$ overall, but z' loses information about the subpopulation A. This negative example highlights the importance of incorporating all the information available (e.g. group membership), not only at the time of decision-making, but also along the way when developing predictors; it serves as yet another rebuke of the approach of "fairness through blindness" [DHP+12].

4 The value of information in fair prediction

In this section, we argue that reasoning about the information content of calibrated predictors provides a lens into understanding how to improve the utility and fairness of predictors, even when the eventual fairness desideratum is based on parity. We discuss a prediction setting based on that of [LDR+18] where a *lender* selects individuals to give loans from a pool of *applicants*. While we use the language of predicting creditworthiness, the setup is generic and can be applied to diverse prediction tasks. [LDR+18] introduced a notion of "delayed impact" of selection policies, which models the potential negative impact on communities of enforcing parity-based fairness as a constraint. We revisit the question of delayed impact as part of a broader investigation of the role of information in fair prediction. We begin with an overview of the prediction setup. Then, we prove our main result: refining the underlying predictions used to choose a selection policy results in an improvement in utility, parity, or impact (or a combination of the three).

4.1 Fair prediction setup

When deciding how to select qualified individuals, the lender's goal is to maximize some expected utility. Specifically, the *utility function* $u : [0,1] \to [-1,1]$ specifies the lender's expected utility from an individual based on their score and a fixed threshold[3] $\tau_u \in [0,1]$ as given in (2). When considering delayed impact, we will measure the expected impact per subpopulation. The *impact function* $\ell : [0,1] \to [-1,1]$ specifies the expected benefit to an individual from receiving a loan based on their score and a fixed threshold τ_ℓ also given in (2).

$$u(p) = p - \tau_u \qquad \ell(p) = p - \tau_\ell \tag{2}$$

[3]Assuming such an affine utility function is equivalent to assuming that the lender receives u_+ from repayments, u_- from defaults, and 0 from individuals that do not receive loans. In this case, the expected utility for score p is $pu_+ + (1-p)u_- = c_u \cdot u(p)$ for some constant c_u. A similar rationale applies to the individuals' impact function.

[LDR+18] models risk-aversion of the lender by assuming that $\tau_u > \tau_\ell$; that is, by choosing accepting individuals with $z(x) \in (\tau_\ell, \tau_u)$, the impact on subpopulations may improve beyond the lender's utility-maximizing policy.

In this setup, we allow the lender to pick a (randomized) group-sensitive selection policy $f : [0,1] \times \{A, B\} \to [0,1]$ that selects individuals on the basis of a predicted score and their sensitive attribute. That is, the selection policy makes decisions about individuals via their score according to some calibrated predictor $z : \mathcal{X} \to [0,1]$ and their sensitive attribute $\mathcal{A} : \mathcal{X} \to \{A, B\}$; for every individual $x \in \mathcal{X}$, the probability that x is selected is given as $f(z(x), \mathcal{A}(x))$.

We will restrict our attention to *threshold policies*; that is, for sensitive attribute A (resp., B), there is some $\tau_A \in [0,1]$, such that $f(v, A)$ is given as $f(v, A) = 1$ if $v > \tau_A$, $f(v, A) = 0$ if $v < \tau_A$ and $f(v, A) = p_A$ for $v = \tau_A$, where $p_A \in [0,1]$ is a probability used to randomly break ties on the threshold. The motivation for focusing on threshold policies is their intuitiveness, widespread use, computational efficiency[4]. The restriction to threshold policies is justified formally in [LDR+18] by the fact that the optimal decision rule in our setting can be specified as a threshold policy under both demographic parity and equalized opportunity.

Given this setup, we can write the expected utility $U^z(f)$ of a policy f based on a calibrated predictor z, that is calibrated on both subpopulations, A and B, as follows.

$$U^z(f) = \sum_{S \in \{A,B\}} \mathop{\mathbf{Pr}}_{x \sim \mathcal{X}} [x \in S] \cdot \left(\sum_{v \in \text{supp}(z)} \mathcal{R}^z_S(v) \cdot f(v, S) \cdot u(v) \right) \tag{3}$$

Recall, $\mathcal{R}^z_S(v) = \mathbf{Pr}_{x \sim S} [z(x) = v]$.

Similarly, the expected impact over the subpopulations $S \in \{A, B\}$ are given as

$$\text{Imp}^z_S(f) = \sum_{v \in \text{supp}(z)} \mathcal{R}^z_S(v) \cdot f(v, S) \cdot \ell(v) \tag{4}$$

Often, it may make sense to constrain the net impact to each group as defined in (4) to be positive, ensuring that the selection policies do not do harm as in [LDR+18].

The following quantities will be of interest to the lender when choosing a selection policy f as a function of z. First, the lender's overall utility $U(f)$ is given as in (3). In the name of fairness, the lender may also be concerned about the disparity of a number of quantities. We will show below that these quantities can be written as a linear function of the selection rule. In particular, for $S \in \{A, B\}$ demographic parity, which serve as our running example, compares the *selection rates* $\beta_S = \mathbf{Pr}_{x \sim S}[x \text{ selected}]$,

$$\beta^z_S(f) = \sum_{v \in \text{supp}(z)} \mathcal{R}^z_S(v) \cdot f(v, S). \tag{5}$$

We may also be concerned about comparing the true positive rates (equalized opportunity) and false positive rates. Recall, $\text{TPR} = \mathbf{Pr}[x \text{ selected} \mid y = 1]$ and $\text{FPR} = \mathbf{Pr}[x \text{ selected} \mid y = 0]$; in

[4]Indeed, without the restriction to threshold policies, many of the *information-theoretic* arguments become easier at the expense of *computational cost*. As z' is a refinement of z, we can always simulate decisions derived from z given z', but in general, we cannot do this efficiently.

this context, we can rewrite these quantities as follows.

$$\mathrm{TPR}_S^z(f) = \frac{1}{r_S} \cdot \sum_{v \in \mathrm{supp}(z)} \mathcal{R}_S^z(v) \cdot f(v, S) \cdot v \tag{6}$$

$$\mathrm{FPR}_S^z(f) = \frac{1}{1 - r_S} \cdot \sum_{v \in \mathrm{supp}(z)} \mathcal{R}_S^z(v) \cdot f(v, S) \cdot (1 - v), \tag{7}$$

where r_S represents the base rate of the subpopulation S; that is, $r_S = \mathbf{Pr}_{(x,y)\sim D_S}[y = 1]$. Another quantity we will track is the positive predictive value, $\mathrm{PPV} = \mathbf{Pr}[y = 1 \mid x \text{ selected}]$.

$$\mathrm{PPV}_S^z(f) = \frac{1}{\beta_S^z(f)} \cdot \left(\sum_{v \in \mathrm{supp}(z)} \mathcal{R}_S^z(v) \cdot f(v, S) \cdot v \right) \tag{8}$$

Note that $\mathrm{PPV}_S^z(f)$ is not a linear function of $f(v, S)$ values, but as we never use positive predictive values directly in the optimizations for choosing a selection policy (or in a parity-based fairness definition), the optimizations are still a linear program. For notational convenience, we may drop the superscript of these quantities when z is clear from the context.

4.2 Refinements in the service of fair prediction

Note that all of the quantities described in Section 4.1 can be written as linear functions of $f(v, S)$. Given a fixed predictor $z : \mathcal{X} \to [0, 1]$, we can expand the quantities of interest; in particular, we note that the linear functions over $f(v, S)$ can be rewritten as linear functions over $z(x)$, where the quantities depend on x only through the predictor z. In this section, we show how refining the predictor used for determining the selection rule can improve the utility, parity, and impact of the optimal selection rule. By the observations above, we can formulate a generic policy-selection problem as a linear program where z controls many coefficients in the program. When we refine z, we show that the value of the program increases. Recalling that different contexts may call for different notions of fairness, we consider a number of different linear programs the lender (or regulator) might choose to optimize. At a high-level, the lender can choose to maximize utility, minimize disparity, or maximize positive impact on groups, while also maintaining some guarantees over the other quantities.

We will consider selection policies given a fixed predictor $z : \mathcal{X} \to [0, 1]$. Note that the parity-based fairness desiderata we consider are of the form $h_A^z(f) = h_B^z(f)$ for some $h \in \{\beta, \mathrm{TPR}, \mathrm{FPR}\}$; rather than requiring equality, we will consider the disparity $|h_A^z(f) - h_B^z(f)|$ and in some cases, constrain it to be less than some constant ε. We also use t_i, t_u to denote lower bounds on the desired impact and utility, respectively. For simplicity's sake, we assume that B is the "protected" group, so we only enforce the positive impact constraint for this group; more generally, we could include an impact constraint for each group. Formally, we consider the following constrained optimizations.

Optimization 1	**Optimization 2**	**Optimization 3**
$\max_f \ U^z(f)$	$\min_f \ \lvert h_A^z(f) - h_B^z(f)\rvert$	$\max_f \ \mathrm{Imp}_B^z(f)$
s.t. $\mathrm{Imp}_B^z(f) \geq t_i$	s.t. $\mathrm{Imp}_B^z(f) \geq t_i$	s.t. $U^z(f) \geq t_u$
$\lvert h_A^z(f) - h_B^z(f)\rvert \leq \varepsilon$	$U^z(f) \geq t_u$	$\lvert h_A^z(f) - h_B^z(f)\rvert \leq \epsilon$
(*Utility Maximization*)	(*Disparity minimization*)	(*Impact Maximization*)

Lemma 4.1. *Let $h \in \{\beta, \mathrm{TPR}, \mathrm{FPR}\}$. Given a calibrated predictor $z : \mathcal{X} \to [0,1]$, Optimization 1,2, and 3 are linear programs in the variables $f(v,S)$ for $v \in \mathrm{supp}(z)$ and $S \in \{A,B\}$. Further, for each program, there is an optimal solution f^* that is a threshold policy.*

We sketch the proof of the lemma. The fact that the optimizations are linear programs follows immediately from the observations that each quantity of interest is a linear function in $f(v,S)$. The proof that there is a threshold policy f^* that achieves the optimal value in each program is similar to the proof of Theorem 4.2 given below. Consider an arbitrary (non-threshold) selection policy f_0; let $h_{S,0} = h_S(f_0)$. The key observation is that for the fixed value of $h_{S,0}$, there is some other threshold policy f where $h_S^z(f) = h_{S,0}$ and $U^z(f) \geq U(f_0)$ and $\mathrm{Imp}_S^z(f) \geq \mathrm{Imp}_S(f_0)$. Leveraging this observation, given any non-threshold optimal selection policy, we can construct a threshold policy, which is also optimal.

We remark that our analysis applies even if considering the more general linear maximization:

Optimization 4

$$\max_f \quad \lambda_U \cdot U^z(f) + \lambda_I \cdot \mathrm{Imp}_B^z(f) - \lambda_\beta \cdot |h_A^z(f) - h_B^z(f)|$$

for any fixed $\lambda_U, \lambda_I, \lambda_\beta \geq 0$.[5] In other words, the arguments hold no matter the relative weighting of the value of utility, disparity, and impact.

Improving the cost of fairness. We argue that in all of these optimizations, increasing information through refinements of the current predictor on both the subpopulations A and B improves this value of the program. We emphasize that this conclusion is true for all of the notions of parity-based fairness we mentioned above. Thus, independent of the exact formulation of fair selection that policy-makers deem appropriate, information content is a key factor in determining the properties of the resulting selection rule. We formalize this statement in the following theorem.

Theorem 4.2. *Let $z, z' : \mathcal{X} \to [0,1]$ be two predictors that are calibrated on disjoint subpopulations $A, B \subseteq \mathcal{X}$. For any of the Optimization 1, 2, 3, 4 and their corresponding fixed parameters, let $OPT(z)$ denote their optimal value under predictor z. If z' refines z on A and B, then $OPT(z') \geq OPT(z)$ for Optimization 1, 3, 4 and $OPT(z') \leq OPT(z)$ for Optimization 2.*

One way to understand Theorem 4.2 is through a "cost of fairness" analysis. Focusing on the utility maximization setting, let U^* be the maximum unconstrained utility achievable by the lender given the optimal predictions p^*. Let $\mathrm{OPT}(z)$ be the optimal value of Optimization 1 using predictions z; that is, the best utility a lender can achieve under a parity-based fairness constraint ($\varepsilon = 0$) and positive impact constraint ($t_i = 0$). If we take the cost of fairness to be the difference between these optimal utilities, $U^* - \mathrm{OPT}(z)$, then Theorem 4.2 says that by refining z to z', *the cost of fairness decreases with increasing informativeness*; that is, $U^* - \mathrm{OPT}(z) \geq U^* - \mathrm{OPT}(z')$. This corollary of Theorem 4.2 corroborates the idea that in some cases the *high perceived cost* associated with requiring fairness might actually be due to the *low informativeness* of the predictions in minority

[5] In particular, each of Optimizations 1, 2, and 3 can be expressed as an instance of Optimization 4 by choosing $\lambda_U, \lambda_I, \lambda_\beta$ to be the optimal dual multipliers for each program. We note that the dual formulation actually gives an alternate way to derive results from [LDR+18]. Their main result can be restated as saying that there exist distributions of scores such that the dual multiplier on the positive impact constraint in Optimization 1 is positve; that is, without this constraint, the utility-maximizing policy will do negative impact to group B.

populations. No matter what the true p^* is, this cost will decrease as we increase information content by refining subpopulations.

For $S \in \{A, B\}$, we use $\mathrm{TPR}_S^z(\beta)$ to denote the true positive rate of the threshold policy with selection rate β for the subpopulation S while using the predictor z[6]. Similarly, $\mathrm{PPV}_S^z(\beta)$, $\mathrm{FPR}_S^z(\beta)$ are defined. The following lemma, which plays a key role in each proof, shows that refinements broadly improve selection policies across these three statistics of interest.

Lemma 4.3. *If z' is a refinement of z on subpopulations A and B, then for $S \in \{A, B\}$, for all $\beta \in [0, 1]$,*

$$\mathrm{TPR}_S^{z'}(\beta) \geq \mathrm{TPR}_S^z(\beta), \qquad \mathrm{FPR}_S^{z'}(\beta) \leq \mathrm{FPR}_S^z(\beta), \qquad \mathrm{PPV}_S^{z'}(\beta) \geq \mathrm{PPV}_S^z(\beta).$$

In particular, the proof of Theorem 4.2 crucially uses the fact that the positive predictive values, true positive rates, and false positive rates improve for *all* selection rates. Leveraging properties of refinements, the improvement across all selection rates guarantees improvement for any fixed objective. As we'll see, the proof actually tells us more: for *any* selection policy using the predictor z, there exists a threshold selection policy that uses the refined predictor z' and *simultaneously* has utility, disparity, and impact that are no worse than under z. In this sense, increasing informativeness of predictors through refinements is an effective strategy for improving selection rules across a wide array of criteria. Still, we emphasize the importance of identifying fairness desiderata and specifying them clearly when optimizing for a selection rule.

For instance, suppose the selection rule is selected by constrained utility maximization with a predictor z and with a refined predictor z'. It is possible that the *optimal* selection policy under the refinement z' will have a lower quantitative impact than the optimal policy under the original predictor z (while still satisfying the impact constraint). If maintaining the impact above a certain threshold is desired, then this should be specified clearly in the optimization used for determining the selection rule. We defer further discussion of these issues to Section 6.

Proofs. Next, we prove Lemma 4.3 and Theorem 4.2.

Proof of Lemma 4.3. Note that for a fixed selection rate β, PPV is maximized by picking the top-most β fraction of the individuals ranked according to p^*, i.e. a threshold policy that selects a β-fraction of the individuals using the Bayes optimal predictor p^*. Similarly, for a fixed selection rate, the TPR and FPR values are also optimized under a threshold selection policy that uses the Bayes optimal predictor p^*.

Recall, we can interpret a refinement as a "candidate" Bayes optimal predictor. In particular, because z' refines z over A and B, we know that z is calibrated not only with respect to the true Bayes optimal predictor p^*, but also with respect to the refinement z' on both subpopulations. Imagining a world in which z' is the Bayes optimal predictor, the PPV, TPR, and FPR must be no worse under a threshold policy derived from z' compared to that of z by the initial observation. Thus, the lemma follows. □

Using Lemma 4.3, we are ready to prove Theorem 4.2.

Proof of Theorem 4.2. Let f be any threshold selection policy under the predictor z. Using f, we will construct a selection policy f' that uses the refined score distribution z' such that

[6]Given a predictor, there is a bijection between selection rates and threshold policies.

where $U^{z'}(f') \geq U^z(f)$, $\text{Imp}_B^{z'}(f') \geq \text{Imp}_B^z(f)$, and $h_A^{z'}(f') = h_A^z(f)$ and $h_B^{z'}(f') = h_B^z(f)$. Here, $h \in \{\beta, \text{TPR}, \text{FPR}\}$ specifies the parity-based fairness definition being used. Thus, taking f to be the optimal solution to any of the Optimizations 1, 2, 3, or 4, we see that f' is a feasible solution to the same optimization and has the same or a better objective value compared to f. Therefore, after optimization, objective values can only get better.

In words, we are saying that refined predictors allow us to get better utility and impact as the original predictor while keeping the parity values the same (for e.g., while keeping the selection rates the same in both subpopulations).

We separately construct f' for each fairness definition (h) as follows:

1. (Demographic Parity) $h = \beta$:

   For $S \in \{A, B\}$, let $\beta_S = \beta_S^z(f)$ be the selection rate of f in the population S. Let f' be the threshold policy that uses the predictor z' and achieves selection rates β_A and β_B in the subpopulations A and B, respectively. By Lemma 4.3, $\text{PPV}_S^{z'}(\beta_S) \geq \text{PPV}_S^z(\beta_S)$ for $S \in \{A, B\}$. The utility of the policy f' can be written as

$$\begin{aligned}
U(f') &= \sum_{S \in \{A,B\}} \Pr_{x \sim \mathcal{X}}[x \in S] \cdot \left(\sum_{v \in \text{supp}(z')} \mathcal{R}_S^{z'}(v) \cdot f'(v, S) \cdot v - \sum_{v \in \text{supp}(z')} \mathcal{R}_S^{z'}(v) \cdot f'(v, S) \cdot \tau_u \right) \\
&= \sum_{S \in \{A,B\}} \Pr_{x \sim \mathcal{X}}[x \in S] \cdot \Big(\beta_S \cdot (\text{PPV}_S^{z'}(\beta_S) - \tau_u) \Big) \\
&\geq \sum_{S \in \{A,B\}} \Pr_{x \sim \mathcal{X}}[x \in S] \cdot (\beta_S \cdot (\text{PPV}_S^{z}(\beta_S) - \tau_u)) \\
&= U(f)
\end{aligned}$$

   Similarly, we can show that the impact on the subpopulation B under f' is at least as good as under f.

2. (Equalized Opportunity) $h = \text{TPR}$:

   Let (β_A, β_B) be the selection rates of policy f on the subpopulations A and B. We know that $\text{TPR}_S^{z'}(\beta_S) \geq \text{TPR}_S^z(\beta_S)$ $(S \in \{A, B\})$ through Lemma 4.3. Let f' be the threshold selection policy corresponding to a selection rates of $\beta'_S, (S \in \{A, B\})$ such that $\text{TPR}_S^{z'}(\beta'_S) = \text{TPR}_S^z(\beta_S)$ $(\leq \text{TPR}_S^{z'}(\beta_S))$. As the true positive rates increase with increasing selection rate, $\beta'_S \leq \beta_S$. The utility of the policy f' can be written as

$$\begin{aligned}
U(f') &= \sum_{S \in \{A,B\}} \Pr_{x \sim \mathcal{X}}[x \in S] \cdot \left(\sum_{v \in \text{supp}(z')} \mathcal{R}_S^{z'}(v) \cdot f'(v, S) \cdot v - \sum_{v \in \text{supp}(z')} \mathcal{R}_S^{z'}(v) \cdot f'(v, S) \cdot \tau_u \right) \\
&= \sum_{S \in \{A,B\}} \Pr_{x \sim \mathcal{X}}[x \in S] \cdot \Big(r_S \cdot \text{TPR}_S^{z'}(\beta'_S) - \beta'_S \cdot \tau_u \Big) \\
&\geq \sum_{S \in \{A,B\}} \Pr_{x \sim \mathcal{X}}[x \in S] \cdot (r_S \cdot \text{TPR}_S^{z}(\beta_S) - \beta_S \cdot \tau_u) \\
&= U(f)
\end{aligned}$$

Similarly, we can show that the impact on the subpopulation B under f' is at least as good as under f.

3. (Equalized False Positive Rate) $h = \text{FPR}$.

   Let (β_A, β_B) be the selection rates of policy f on the subpopulations A and B. We know that $\text{FPR}_S^{z'}(\beta_S) \leq \text{FPR}_S^{z}(\beta_S)$ $(S \in \{A, B\})$ through Lemma 4.3. Let f' be the threshold selection policy corresponding to a selection rates of $\beta'_S, (S \in \{A, B\})$ such that $\text{FPR}_S^{z'}(\beta'_S) = \text{FPR}_S^{z}(\beta_S)$ $(\geq \text{FPR}_S^{z'}(\beta_S))$. As the false postive rates increase with increasing selection rate, $\beta'_S \geq \beta_S$. The utility of the policy f' can be written as

$$\begin{aligned}
U(f') &= \sum_{S \in \{A,B\}} \mathop{\mathbf{Pr}}_{x \sim \mathcal{X}}[x \in S] \cdot \left(\sum_{v \in \text{supp}(z')} \mathcal{R}_S^{z'}(v) \cdot f'(v, S) \cdot v - \sum_{v \in \text{supp}(z')} \mathcal{R}_S^{z'}(v) \cdot f'(v, S) \cdot \tau_u \right) \\
&= \sum_{S \in \{A,B\}} \mathop{\mathbf{Pr}}_{x \sim \mathcal{X}}[x \in S] \cdot \left(\beta'_S - (1 - r_S) \cdot \text{FPR}_S^{z'}(\beta'_S) - \beta'_S \cdot \tau_u \right) \\
&= \sum_{S \in \{A,B\}} \mathop{\mathbf{Pr}}_{x \sim \mathcal{X}}[x \in S] \cdot \left(\beta'_S \cdot (1 - \tau_u) - (1 - r_S) \cdot \text{FPR}_S^{z'}(\beta'_S) \right) \\
&\geq \sum_{S \in \{A,B\}} \mathop{\mathbf{Pr}}_{x \sim \mathcal{X}}[x \in S] \cdot (\beta_S \cdot (1 - \tau_u) - (1 - r_S) \cdot \text{FPR}_S^{z}(\beta_S)) \\
&= U(f)
\end{aligned}$$

   Similarly, we can show that the impact on the subpopulation B under f' is at least as good as under f.

This completes the proof of the theorem. □

5 A mechanism for refining predictors

In this section, we outline a mechanism for obtaining refinements of predictors. We start by describing an algorithm, `merge`, which given two calibrated predictors, produces a new refined predictor that incorporates the information from both predictors (in a sense we make formal). For notational convenience, we describe how to refine a predictor over $\mathcal{X}$. The arguments here extend easily to refining over a partition of $\mathcal{X}$ by refining each part separately. We discuss the generality of the approach at the end of the section and elaborate on the possibility of refinements on overlapping subpopulations briefly in Section 6.

Given a predictor z, we can evaluate the information content $I(z)$ directly; estimating the information loss $L(p^*; z)$, however, is generally impossible. Indeed, without assumptions on the structure of p^* or the ability to sample every individual's outcome repeatedly (and independently), we cannot reason about information-theoretic quantities like $I(p^*)$. Still, supposing that the information loss of z is sufficiently large, we would like to be able to certify this fact and ideally, bring the loss down.

The most obvious way to demonstrate that a predictor z can be refined would be to exhibit some calibrated $q : \mathcal{X} \to [0, 1]$ such that $I(q) > I(z)$. That said, expecting that we could obtain such a

q seems to sidestep the question of how to update a predictor to improve its information content. Even if we were able to obtain some q where $I(q) > I(z)$, it is not clear that q would be a "better" predictor. In particular, q might contain *different* information than z; recall that such examples motivated the definition of a refinement in the first place. Still, intuitively, if q is not a refinement of z and contains very different information than z, then q should be useful in identifying ways to improve the informativeness of z. Further, this intuition does not seem to rely on the fact that $I(q) > I(z)$; as long as q contains information that isn't "known" to z, then incorporating the information into z should reduce the information loss.

To formalize this line of reasoning, first, we need to make precise what we mean when we say that q is far from refining z. Recalling the definition of a refinement, consider the logical negation of the statement that "q refines z."

$$\neg \left(\forall v \in \text{supp}(z) : \mathop{\mathbf{E}}_{x \sim \mathcal{X}_{z(x)=v}} [q(x)] = v \right) \iff \exists v \in \text{supp}(z) : \mathop{\mathbf{E}}_{x \sim \mathcal{X}_{z(x)=v}} [q(x)] \neq v.$$

Extending this logical formulation, we define the following divergence to capture quantitatively how far q is from refining z.

Definition (Refinement distance)**.** *Let $q, z : \mathcal{X} \to [0,1]$ be calibrated predictors. The* refinement distance *from z to q is given as*

$$D_R(z; q) = \sum_{v \in \text{supp}(z)} \mathcal{R}^z(v) \cdot \left| \mathop{\mathbf{E}}_{x \sim \mathcal{X}_{z(x)=v}} [q(x)] - v \right|.$$

Note that $D_R(z; q)$ is not symmetric; in particular, if q refines z and contains more information $I(q) > I(z)$, then $D_R(z; q) = 0$, but $D_R(q; z) > 0$. Intuitively, the refinement distance averages the refinement "disagreements" over all values in the support. We show that, under calibration, these disagreements can be reconciled to improve the overall information content. With the notion of refinement distance in place, we can state the main algorithmic result – a simple algorithm for aggregating the information of multiple calibrated predictors into a single calibrated predictor.

Theorem 5.1. *Given two calibrated predictors $q, z : \mathcal{X} \to [0,1]$, Algorithm 1 produces a new calibrated predictor $\rho : \mathcal{X} \to [0,1]$ such that ρ is a refinement of both z and q. Further, $I(\rho) > \max\left\{ I(z) + 4 \cdot D_R(q; z)^2, I(q) + 4 \cdot D_R(z; q)^2 \right\}$.*

We state the theorem generally, making no assumptions about $D_R(q; z)$ or $D_R(z; q)$. In particular, as we alluded to earlier, if q already refines z, then $D_R(z; q) = 0$, so there will be no information gain. Algorithm 1, which we refer to as `merge`, describes the procedure. We describe the algorithm in the statistical query model, where we assume query access to aggregate statistics about $p^*(x)$. At the end of the section, we discuss the sample complexity needed to answer such statistical queries accurately. The `merge` algorithm builds a new calibrated predictor ρ from q and z by considering the set of individuals who receive $q(x) = u$ and $z(x) = v$ for each $u \in \text{supp}(q)$ and $v \in \text{supp}(z)$. For each of these sets, the merged predictor adjusts the prediction to have the correct expectation. The proof of Theorem 5.1 follows from a standard potential function analysis; further, the sample complexity needed to answer the statistical queries accurately is bounded.

We break the proof of Theorem 5.1 into two lemmas. First, note that the `merge` procedure is symmetric with respect to q and z. Thus, any statements about the output ρ in terms of one of the inputs z will also be true with respect to the input q.

Algorithm 1: `merge(z,q)`

Given: $z, q : \mathcal{X} \to [0,1]$ calibrated predictors
Output: $\rho : \mathcal{X} \to [0,1]$ a refinement of z and q

- Let $\mathcal{Z} = \left\{\mathcal{X}_{z(x)=v} : v \in \text{supp}(z)\right\}$
- Let $\mathcal{Q} = \left\{\mathcal{X}_{q(x)=u} : u \in \text{supp}(q)\right\}$
- For $Z_v \in \mathcal{Z}$ and $Q_u \in \mathcal{Q}$:
  - $\mathcal{X}_{vu} = Z_v \cap Q_u$
  - $\rho(x) \leftarrow \mathop{\mathbf{E}}_{x \sim \mathcal{X}_{vu}} [p^*(x)]$

Lemma 5.2. *Let ρ be the output of Algorithm 1 on two calibrated predictors $z, q : \mathcal{X} \to [0,1]$ as input. ρ is a refinement of z.*

Proof. We use the notation established in Algorithm 1. In particular, we refer to the conditional score distribution $\mathcal{R}^q_{Z_v}$ where $\mathcal{R}^q_{Z_v}(u) = \mathbf{Pr}_{x \sim \mathcal{X}}\left[q(x) = u \;\middle|\; z(x) = v\right]$. Consider the expectation of ρ over the level sets of z, $\{Z_v : v \in \text{supp}(z)\}$.

$$\begin{aligned}
\mathop{\mathbf{E}}_{x \sim Z_v}[\rho(x)] &= \sum_{u \in \text{supp}(q)} \mathcal{R}^q_{Z_v}(u) \cdot \mathop{\mathbf{E}}_{x \sim \mathcal{X}_{vu}}[\rho(x)] \\
&= \sum_{u \in \text{supp}(q)} \mathcal{R}^q_{Z_v}(u) \cdot \mathop{\mathbf{E}}_{x \sim \mathcal{X}_{vu}}\left[\mathop{\mathbf{E}}_{x \sim \mathcal{X}_{vu}}[p^*(x)]\right] && (9) \\
&= \sum_{u \in \text{supp}(q)} \mathcal{R}^q_{Z_v}(u) \cdot \mathop{\mathbf{E}}_{x \sim Z_v}\left[p^*(x) \;\middle|\; q(x) = u\right] && (10) \\
&= \mathop{\mathbf{E}}_{x \sim Z_v}[p^*(x)]
\end{aligned}$$

where (9) follows by the assignment rule of $\rho(x)$ for $x \in \mathcal{X}_{vu}$; and (10) follows from exploiting $\mathcal{X}_{vu} = \{x \in Z_v : q(x) = u\}$. By the calibration of z, we see the final expression is equal to v. This argument is independent of v, so for all $v \in \text{supp}(z)$, $\mathbf{E}_{x \sim Z_v}[\rho(x)] = v$; thus, by definition, ρ refines z. □

The next lemma shows that the information of ρ increases based on the refinement distance. Note that in combination Lemma 5.2 and Lemma 5.3 prove Theorem 5.1.

Lemma 5.3. *Let ρ be the output of Algorithm 1 on two calibrated predictors $z, q : \mathcal{X} \to [0,1]$ as input. Then,*

$$I(\rho) \geq I(z) + 4 \cdot D_R(q; z)^2.$$

Proof. We can lower bound the resulting information content of ρ by reasoning about the difference $I(\rho) - I(z)$. Note that by Lemma 5.2, ρ is a refinement of z; further, by Proposition 3.2, we can

express $I(\rho) - I(z)$ as $L(\rho; z)$. Expanding the information loss, we can lower bound the gain in information, which shows the lemma.

$$\begin{aligned}
\frac{1}{4} \cdot L(\rho; z) &= \mathop{\mathbf{E}}_{x \sim \mathcal{X}} \left[(\rho(x) - z(x))^2 \right] \\
&= \sum_{u \in \text{supp}(q)} \mathcal{R}^q(u) \cdot \mathop{\mathbf{E}}_{x \sim Q_u} \left[(z(x) - \rho(x))^2 \right] \\
&\geq \sum_{u \in \text{supp}(q)} \mathcal{R}^q(u) \cdot \left(\mathop{\mathbf{E}}_{x \sim Q_u} \left[z(x) - \rho(x) \right] \right)^2 && (11) \\
&= \sum_{u \in \text{supp}(q)} \mathcal{R}^q(u) \cdot \left(\mathop{\mathbf{E}}_{x \sim Q_u} \left[z(x) \right] - u \right)^2 && (12) \\
&\geq \left(\sum_{u \in \text{supp}(q)} \mathcal{R}^q(u) \cdot \left| \mathop{\mathbf{E}}_{x \sim Q_u} \left[z(x) \right] - u \right| \right)^2 && (13) \\
&\geq D_R(q; z)^2 && (14)
\end{aligned}$$

where (11) follows by Jensen's Inequality; (12) notes that $\mathbf{E}_{x \sim Q_u}[\rho(x)] = u$ because q is calibrated and ρ refines q; (13) applies Jensen's inequality again; and (14) follows by the definition of refinement distance. □

One appealing consequence of Theorem 5.1 is that the number of times a predictor needs to be significantly updated is bounded. In particular, suppose we are merging two calibrated predictors q, z; for any $\eta \geq 0$, we'll say the operation is an η-merge if $\min \{D_R(z; q), D_R(q; z)\} \geq \eta$. In this case, the information content of the result predictor will increase by at least $\Omega(\eta^2)$ and any given predictor can be η-merged at most $O(1/\eta^2)$ times. In other words, as long as the information being combined is not too similar, then the number of such merge updates is bounded.

Interpreting the updates. As described, the `merge` algorithm takes two different calibrated predictors and combines them into a refinement. In settings where the lender wants to combine predictions from different sources, this algorithmic model is naturally well-motivated. Still, there are other settings where the `merge` algorithm can be applied. One natural way we can specify new information content is by giving the predictor an additional feature. Specifically, consider some a boolean feature $\phi : \mathcal{X} \to \{0, 1\}$. We define the predictor $q_\phi : \mathcal{X} \to [0, 1]$ to be $q_\phi(x) = \mathbf{E}_{x' \sim \mathcal{X}} \left[p^*(x') \;\middle|\; \phi(x') = \phi(x) \right]$. This predictor gives the expected value over the set of individuals where $\phi(x) = 1$ (resp., $\phi(x) = 0$); thus, the predictor is calibrated. Merging z with q_ϕ incorporates the information in the boolean feature ϕ into the predictions of z. In particular, the information content framework gives us a way to reason about the marginal informativeness of individual boolean features; the greater the difference between $\mathbf{E}_{x \sim \phi^{-1}(0)}[p^*(x)]$ and $\mathbf{E}_{x \sim \phi^{-1}(1)}[p^*(x)]$, the more informative.

This perspective is particularly salient when we think external regulation of predictors. For example, consider some subpopulation $S \subseteq \mathcal{X}$. One way to provide evidence that S is experiencing discrimination under z would be to demonstrate that merging the predictor q_S into z significantly changes the information content. This could occur because the quality of individuals in S are

consistently underestimated by z or because the quality of individuals in $\mathcal{X} \setminus S$ are consistently overestimated.

Implementing the merge from samples. While we presented the `merge` algorithm assuming access to a statistical query oracle, in practice, we want to estimate the necessary statistical queries from data. We assume that the predictions are discretized to precision α; that is, we represent the interval $[0, 1]$ as $[\alpha/2, 3\alpha/2, \ldots, 1 - \alpha/2]$. We argue that the number of samples needed to obtain accurate statistics in this model is bounded as follows.

Proposition 5.4. *Consider an execution of Algorithm 1 such that* $\min_{\mathcal{X}_{vu}} \mathbf{Pr}_{x \sim \mathcal{X}} [x \in \mathcal{X}_{vu}] \geq \gamma$. *Then from* $m \geq \tilde{\Omega}\left(\frac{\log(1/\delta)}{\gamma\alpha^2}\right)$ *random samples from* $\mathcal{D}$, *with probability* $1 - \delta$, *every statistical query can be answered with some* q_{vu} *such that* $|q_{vu} - \mathbf{E}_{x \sim \mathcal{X}_{vu}} [p^*(x)]| < \alpha/2$.

Proof. The argument follows by a standard uniform convergence argument. To start, note that there are at most $1/\alpha^2$ queries to answer. Suppose we have t random samples $(x_i, y_i) \sim \mathcal{D}$ conditioned on $x_i \in \mathcal{X}_{vu}$ for all $i \in [t]$. Let q_{vu} denote the empirical expectation of y_i's on these t samples over $\mathcal{X}_{vu}$; that is,

$$q_{vu} \triangleq \frac{1}{t} \sum_{i=1}^{t} y_i.$$

By Hoeffding's inequality:

$$\mathbf{Pr}\left[\left| q_{vu} - \mathop{\mathbf{E}}_{x \sim \mathcal{X}_{vu}} [y \mid x] \right| > \alpha \right] \leq 2e^{2t\alpha^2}.$$

If $t \geq \Omega\left(\frac{\log(2/\delta\alpha^2)}{\alpha^2}\right)$, then the probability of failure is at most $\alpha^2\delta/2$. Union bounding over the queries, the probability of failure is at most $\delta/2$.

Thus, we need to bound the sample complexity needed to hit each $\mathcal{X}_{vu}$ at least t times. By assumption, each $\mathcal{X}_{vu}$ has density at least γ. Thus, for each $\mathcal{X}_{vu}$, the probability that a random sample from $(x, y) \sim \mathcal{D}$ has $x \in \mathcal{X}_{vu}$ is at least γ. If we take $\Omega(\log(2t/\delta)/\gamma)$ such samples, then the probability every sample misses $\mathcal{X}_{vu}$ is at most $\delta/2t$. Thus, if we take $\Omega(t \log(2t/\delta)/\gamma)$ samples, each $\mathcal{X}_{vu}$ will have at least t samples with probability at least $1 - \delta/2$.

By union bound, the proposition follows. □

6 Discussion

In this work, we identify information disparity as a potential source of discrimination in prediction tasks. We provide an introduction to key concepts of information content and loss, and show how improving the information content of predictions improves the resulting fairness of the downstream decisions. In particular, our results show when a lender does not have sufficient statistical or computational resources to learn a predictor that achieves small squared error across all significant subpopulations, issues of unfairness may arise due to differential information loss.

The information content of a predictor z can be significantly larger on the majority population S than the minority T for a number of reasons.

- Despite optimal predictions, the individuals in S are inherently more predictable than those in T; i.e. $z \approx p^*$ and $I_S(p^*) > I_T(p^*)$. If this (controversial) hypothesis is true, there may be no way to improve the predictions further, and some degree of disparity may be unavoidable. Note that in general this condition cannot be verified from data. Still, the assumption that $I(z) = I(p^*)$ can be falsified by finding a way to give more informative predictions.
- Nontrivial information loss has occurred in z on T compared to S; i.e. $L_S(p^*; z) < L_T(p^*; z)$. Such information loss could result from purely information theoretic issues (features used for prediction are not sufficiently expressive in T), a mix of informational and computational issues (not enough data from the minority to learn a predictor from a sufficiently rich hypothesis class), or purely computational issues (suboptimal learning in T due to optimization for $S \cup T$). Each source of information disparity has a different own solution (collecting better features, collecting more data, re-training with awareness of the population T, respectively), but the tools we present for reasoning about information content apply broadly.

As such, improving the information content of predictions may require collecting additional features or data. Once collected, the `merge` procedure provides a relatively inexpensive way of incorporating new information into the predictions while retaining certain "quality" of the selection rule. In practice, when the sensitive group T is known, it may make sense to simply retrain the prediction model with awareness of T.

Overlapping subpopulations. The proposed `merge` algorithm provides a simple and efficient approach for producing refinements in applications where the sensitive populations are well understood. Often, as highlighted in [HKRR18, KNRW18, KRR18], the subpopulations in need of protection may be hard to anticipate. These recent works have studied notions of *multi-fairness* that aim to strengthen notions of group fairness by enforcing statistical constraints not just overall, but on a rich family of subgroups. In particular, [HKRR18] introduces a notion called *multicalibration*, which informally, guarantees calibration across all subpopulations specified by a given set system $\mathcal{C}$. We observe that the guarantees of multicalibration can be reinterpreted in the language of refinements: a predictor that is multicalibrated with respect to $\mathcal{C}$ is simultaneously a refinement for all $q_c : \mathcal{X} \to [0, 1]$ where $q_c(x) = \mathbf{E}_{x' \sim \mathcal{X}} \left[p^*(x') \;\middle|\; c(x') = c(x) \right]$ for every $c \in \mathcal{C}$. In this sense, multicalibration may be an effective approach to improving information across subpopulations, when the groups that might be experiencing discrimination are unknown or overlapping. An interesting question is whether some of the analysis in the present work can be applied to understand better the connections between multicalibration and the work of [KNRW18] which studies the notion of rich subgroup fairness under demographic parity and equalized opportunity.

Choosing fairness constraints. Understanding precisely the guarantees of the specified fairness constraints is particularly important for interpreting the results of Section 4. In particular, refining predictions is guaranteed to improve the value of the *stated program*. We emphasize the importance of faithfully translating fairness desiderata into mathematical requirements.

For instance, suppose policy-makers want to increase representation for historically-disenfranchised populations. An appealing – but misguided – translation of this goal would require demographic parity; intuitively, if the lender is required to have equalized selection rates across groups, they might increase the selection rate in the minority to match that of the majority. Still, demographic

parity only requires parity of selection rates which could also be achieved by reducing the selection in the majority. Further, refinements under demographic parity constraints, might cause the selection rates in the minority to *decrease*. By increasing information, we might uncover that fewer individuals are actually above the tolerable risk than the previous predictions suggested; as such, fewer individuals might be deemed qualified for a loan.

The framework proposed in Section 4 is compatible with a variety of constraints and objectives, including explicitly lower bounding the group selection rates. Thus, increasing the selection rate in a given population can always be achieved by directly constraining the selection rule. An appealing aspect of the framework is that it allows policy-makers to experiment with constraints and objectives to understand the downstream effects of their policies, given the current set of predictions. For instance, policy-makers can evaluate how lower bounding the selection rate in a group will affect the impact of the policy on this group. Such experimentation with the programs from Section 4 may help to guide future policy decisions.

Changing environments. The present work focuses on a setting where the true risk scores of the underlying population does not change; that is, we assumed that the Bayes optimal predictor p^* remains fixed while producing better and better refinements. In real life, the true risk of individuals may change as their environment changes, and possibly even *as a result of the prior decisions made by the lender*, as suggested by [LDR$^+$18]. An exciting direction for future investigation would study a setting of dynamic p^*, with the goal of ensuring long-term fairness and impact. A specific challenge is finding an efficient (in terms of sample and time complexity) procedure for maintaining calibration when p^* changes over time. Further, we showed the importance of increasing informativeness of predictors for underrepresented populations, but required access to random samples from this population. In settings where random exploration may cause harm to uncertain populations (e.g. by raising the default rate) how can we improve information without causing the inherent capabilities of sensitive subpopulations to deteriorate?

Conclusion. We reiterate that the validity of every notion of fairness rests on some set of assumptions. Many approaches to fair classification assume implicitly that the predicted risk scores represent the true risk scores. Predicted risk scores are the result of an extensive pipeline of data collection and computational modeling; when data is limited for minority populations and modeling is focused on fidelity in the majority populations, the resulting predictions may not be appropriately informative in the minority. In the case that the differences arise because of suboptimal predictions, increasing information through refinements provide a simple but effective approach for improving the utility, fairness, and impact of the decision-maker's policy.

Acknowledgments. The authors thank Cynthia Dwork and Guy N. Rothblum for many helpful conversations throughout the development of this work. We thank Moritz Hardt, Gal Yona, and anonymous reviewers for feedback on earlier versions of the work.

References

[ALMK16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016.

[BCZ+16] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Neural Information Processing Systems*, 2016.

[BG18] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT**, 2018.

[Bla53] David Blackwell. Equivalent comparisons of experiments. *The annals of mathematical statistics*, 1953.

[Bri50] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 1950.

[CG18] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint 1808.00023*, 2018.

[Cho17] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 2017.

[CJS18] Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *Neural Information Processing Systems*, 2018.

[Cré82] Jacques Crémer. A simple proof of blackwell's "comparison of experiments" theorem. *Journal of Economic Theory*, 1982.

[DF81] Morris H DeGroot and Stephen E Fienberg. Assessing probability assessors: Calibration and refinement. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF STATISTICS, 1981.

[DHP+12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In *ITCS*, 2012.

[FK18] Irfan Faizullabhoy and Aleksandra Korolova. Facebook's advertising platform: New attack vectors and the need for interventions. *arXiv preprint 1803.10099*, 2018.

[FV98] Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 1998.

[GBR07] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2007.

[GR07] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 2007.

[HKRR18] Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Calibration for the (computationally-identifiable) masses. *ICML*, 2018.

[HM19] Ben Hutchinson and Margaret Mitchell. 50 years of testing (un)fairness: Lessons for machine learning. In *FAT**, 2019.

[HPS16] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Neural Information Processing Systems*, 2016.

[ILZ19] Nicole Immorlica, Katrina Ligett, and Juba Ziani. Access to population-level signaling as a source of inequality. *FAT**, 2019.

[JKMR16] Matthew Joseph, Michael Kearns, Jamie H. Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Neural Information Processing Systems*, 2016.

[KLMR18] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic fairness. In *AEA Papers and Proceedings*, 2018.

[KMR17] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *ITCS*, 2017.

[KNRW18] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *ICML*, 2018.

[KRR18] Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Fairness through computationally-bounded awareness. *Neural Information Processing Systems*, 2018.

[KRZ19] Sampath Kannan, Aaron Roth, and Juba Ziani. Downstream effects of affirmative action. *FAT**, 2019.

[LDR+18] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *ICML*, 2018.

[MPB18] Shira Mitchell, Eric Potash, and Solon Barocas. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint 1811.07867*, 2018.

[PRW+17] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon M. Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. In *Neural Information Processing Systems*, 2017.

[SbF+19] Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *FAT**, 2019.

[STM+18] Samira Samadi, Uthaipon Tantipongpipat, Jamie H Morgenstern, Mohit Singh, and Santosh Vempala. The price of fair pca: One extra dimension. In *Neural Information Processing Systems*, 2018.

A Measuring information through Shannon entropy

For completeness, we briefly discuss how to relate the notions of information defined in Section 3 to an analogous notion of information, defined through Shannon entropy. In particular, rather than defining information content in terms of the variance of y given $z(x)$, we could have defined it in terms of the Shannon entropy of this random variable. In particular, the entropy of a Bernoulli random variable with expectation p is captured by the binary entropy function $H_2(p)$ where

$$H_2(p) = -p \cdot \log(p) - (1-p) \cdot \log(1-p).$$

For a calibrated predictor $z : \mathcal{X} \to [0,1]$, let $I_S^{\text{ent}}(z) = 1 - \mathbf{E}_{x \sim S}[H_2(x)]$ denote the *entropic information content*, parameterized by the binary entropy (rather than variance). The binary entropy $H_2(p)$ always upper bounds the scaled variance $4 \cdot p(1-p)$, with equality at $p \in \{0, 1/2, 1\}$. As a consequence, the following inequality holds.

$$I_S^{\text{ent}}(z) \leq I_S(z)$$

When information is parameterized by entropy, the corresponding natural notion of "information loss" is parameterized by the expected KL-divergence. Specifically, denote by $D_{KL}(p;q)$ the KL-divergence between two Bernoulli distributions with expectations p and q, respectively, defined as

$$D_{KL}(p;q) = p \cdot \log\left(\frac{p}{q}\right) + (1-p) \cdot \log\left(\frac{1-p}{1-q}\right).$$

Again, for a calibrated predictor $z : \mathcal{X} \to [0,1]$, let $L_S^{\text{ent}}(p^*;z) = \mathbf{E}_{x \sim S}[D_{KL}(p^*(x); q(x))]$ denote the *entropic information loss*. We can relate the information loss (based on squared error) to the entropic information loss (based on KL-divergence).

Proposition A.1. *For a calibrated predictor $z : \mathcal{X} \to [0,1]$, where $\text{supp}(z) \subseteq \{0,1\} \cup [\alpha, 1-\alpha]$ for some constant $\alpha > 0$, the entropic information loss is within a constant factor of the information loss.*

$$L_S(p^*;z) \leq 2\ln(2) \cdot L_S^{ent}(p^*;z) \leq \frac{1}{\alpha} \cdot L_S(p^*;z)$$

Proposition A.1 is a direct corollary of Pinsker's inequality, which relates the KL-divergence to the statistical distance between two probability distributions. As in Proposition 3.1 we can show that this entropic information loss can be expressed as a difference in entropic information content for calibrated predictors.

Proposition A.2. *Let $p^* : \mathcal{X} \to [0,1]$ denote the Bayes optimal predictor. Suppose for $S \subseteq \mathcal{X}$, $z : \mathcal{X} \to [0,1]$ is calibrated on S. Then,*

$$L_S^{ent}(p^*;z) = I^{ent}(p^*) - I^{ent}(z).$$

Proof. Again, we expand the entropic information loss and rearrange.

$$\begin{aligned}
\mathop{\mathbf{E}}_{x \sim S}[D_{KL}(p^*(x); z(x))] &= \mathop{\mathbf{E}}_{x \sim S}\left[p^*(x) \cdot \log\left(\frac{p^*(x)}{z(x)}\right) + (1-p^*(x)) \cdot \log\left(\frac{p^*(x)}{1-z(x)}\right)\right] \\
&= \mathop{\mathbf{E}}_{x \sim S}\left[p^*(x) \cdot \log\left(\frac{1}{z(x)}\right) + (1-p^*(x)) \cdot \log\left(\frac{1}{1-z(x)}\right)\right] - \mathop{\mathbf{E}}_{x \sim S}[H_2(p^*(x))]
\end{aligned}$$

Leveraging the fact that z is calibrated, we expand the first term as follows.

$$\begin{aligned}
&\mathop{\mathbf{E}}_{x\sim S}\left[p^*(x)\cdot\log\left(\frac{1}{z(x)}\right)+(1-p^*(x))\cdot\log\left(\frac{1}{1-z(x)}\right)\right]\\
&=\sum_{v\in\mathrm{supp}(z)}\mathcal{R}^z_S(v)\cdot\mathop{\mathbf{E}}_{x\sim S_v}\left[p^*(x)\cdot\log\left(\frac{1}{\mathbf{E}_{x'\sim S_v}[p^*(x')]}\right)+(1-p^*(x))\cdot\log\left(\frac{1}{1-\mathbf{E}_{x'\sim S_v}[p^*(x')]}\right)\right]\\
&=\sum_{v\in\mathrm{supp}(z)}\mathcal{R}^z_S(v)\cdot\left(\mathop{\mathbf{E}}_{x\sim S_v}[p^*(x)]\cdot\log\left(\frac{1}{\mathbf{E}_{x\sim S_v}[p^*(x)]}\right)+(1-\mathop{\mathbf{E}}_{x\sim S_v}[p^*(x)])\cdot\log\left(\frac{1}{1-\mathbf{E}_{x\sim S_v}[p^*(x)]}\right)\right)\\
&=\sum_{v\in\mathrm{supp}(z)}\mathcal{R}^z_S(v)\cdot H_2\left(\mathop{\mathbf{E}}_{x\sim S_v}[p^*(x)]\right)\\
&=\sum_{v\in\mathrm{supp}(z)}\mathcal{R}^z_S(v)\cdot H_2(v)\\
&=\mathop{\mathbf{E}}_{x\sim S}[H_2(z(x))]
\end{aligned}$$

Thus, combining the equlaities, we see that

$$L^{\mathrm{ent}}_S(p^*;z)=\mathop{\mathbf{E}}_{x\sim S}[H_2(z(x))]-\mathop{\mathbf{E}}_{x\sim S}[H_2(p^*(x))]=I^{\mathrm{ent}}_S(p^*)-I^{\mathrm{ent}}_S(z).$$

□

As a final note, we observe that the entropic information content of a calibrated predictor can be related to the expected log-likelihood of the predictor. Specifically, given a collection of labeled data $(x_1,y_1),\ldots,(x_m,y_m)$ where $y_i\sim\mathrm{Ber}(p^*(x_i))$, the likelihood function $\mathcal{L}(z;\{(x_i,y_i)\})$ is given as follows.

$$\mathcal{L}(z;\{(x_i,y_i)\})=\prod_{i=1}^m z(x_i)^{y_i}\cdot(1-z(x_i))^{1-y_i}$$

As such, we say the normalized log-likelihood $\ell(z;\{(x_i,y_i)\})$ is given as

$$\ell(z;\{(x_i,y_i)\})=\frac{1}{m}\sum_{i=1}^m y_i\cdot\log(z(x_i))+(1-y_i)\cdot\log(1-z(x_i)).$$

Suppose the samples $(x_1,y_1),\ldots,(x_m,y_m)$ are drawn such that each $x_i\in S$ for some subset $S\subseteq\mathcal{X}$, then the expected log-likelihood of z is given as follows.

$$\begin{aligned}
\mathop{\mathbf{E}}_{\substack{x_i\sim S\\ y_i\sim\mathrm{Ber}(p^*(x))}}[\ell(z;\{(x_i,y_i)\})]&=\mathop{\mathbf{E}}_{\substack{x\sim S\\ y\sim\mathrm{Ber}(p^*(x))}}[y\cdot\log(z(x))+(1-y)\cdot\log(1-z(x))]\\
&=\mathop{\mathbf{E}}_{x\sim S}[p^*(x)\cdot\log(z(x))+(1-p^*(x))\cdot\log(1-z(x))]\\
&=-\mathop{\mathbf{E}}_{x\sim S}[H_2(z(x))] \qquad (15)\\
&=I^{\mathrm{ent}}_S(z)-1
\end{aligned}$$

where (15) follows from the analysis above given in the proof of Proposition A.2. In other words, as we increase the information content of a calibrated predictor, in expectation, the likelihood of the predictor increases (in expectation over a fresh sample of data). At the extreme, the calibrated predictor that maximizes the expected likelihood is p^*.