Fairness Beyond Non-discrimination: Feature Selection for Fair Decision Making

Nina Grgić-Hlača†, Muhammad Bilal Zafar‡, Krishna P. Gummadi†, Adrian Weller‡

†Max Planck Institute for Software Systems (MPI-SWS), Germany
‡University of Cambridge, UK

[nghla, mzafer, gumadi]@mpi-sws.org, adrian.weller@eng.cam.ac.uk

ABSTRACT

Machine learning methods are increasingly being used to inform, or sometimes even directly to make, important decisions about humans. A number of recent works have focused on the fairness of the outcomes of such decisions, particularly on avoiding decisions that affect users of different sensitive groups (e.g., race, gender) disparately. In this paper, we propose to consider the fairness of the process of decision making. Process fairness can be measured by estimating the degree to which people consider various features to be fair to use when making an important decision. We examine two tasks to predict criminal risk: recidivism risk prediction in the dataset considered by ProPublica relating to the COMPAS system, and prediction of illegal weapon possession in the stop-question-and-frisk dataset provided by New York Police Department. We introduce new measures of people’s discomfort with using various features, show how these measures can be estimated, and consider the effect of removing the uncomfortable features on prediction accuracy and on outcome fairness. We provide a fast submodular optimization mechanism to control the tradeoff between process fairness and prediction accuracy. Our empirical findings suggest that process fairness may be achieved with little cost to outcome fairness, but that some loss of accuracy is unavoidable.

1 INTRODUCTION

As machine learning methods are increasingly being used in decision making scenarios that affect human lives (such as credit risk assessments and recidivism risk prediction), there is a growing concern about the fairness of such decision making. These concerns have spawned much recent research to find methods for detecting and avoiding unfairness in decision making [11, 13, 19, 22, 24, 27, 28]. In this work, we revisit the foundational notions of fairness that underlie these fair learning and unfairness detection methods.

We argue that the notions of fairness underlying much of the prior work are centered around the outcomes of the decision process. They are inspired in large part by the application of antidiscrimination laws in various countries [1], under which decision policies or practices (implemented by humans) can be declared as discriminatory based on their effects on people belonging to certain sensitive demographic groups (e.g., gender, race). For instance, the notions of “individual fairness” in [11], “situational testing” in [22], and “disparate treatment” in [27], consider individuals who belong to different sensitive groups, yet share similar non-sensitive features (qualifications), and require them to receive similar decision outcomes. Similarly, the notions of “group fairness” in [28] and “disparate impact” in [27] are based on different sensitive groups (e.g., males and females) receiving beneficial decision outcomes in similar proportions. Finally, the notion of “disparate mistreatment” in [26] is rooted in the desire for different sensitive demographic groups to experience similar rates of errors in decision outcomes. Thus, in prior works, the fairness of decision making has been evaluated based on the decision outcomes.

In this paper, we make the case for notions of fairness that are based on the process of decision making rather than on the outcomes. Our notions of process fairness are motivated by the observation that in many decision making scenarios, humans tend to have a moral sense for whether or not it is fair to use an input feature in the decision making process. For instance, consider the task of predicting recidivism risk for an offender. COMPAS is a commercial recidivism prediction tool that relies on a number of different types of user features, such as information about Criminal history, Family criminality, Work and Social environment of the offender. In a user survey that we conducted, we found that a strong majority of users felt that it was fair to use Criminal history, but unfair to use Family criminality. On the other hand, features Work and Social environment were deemed as fair and unfair (respectively) by only a weak majority of users.

Such societal consensus (strong or weak) on the fairness of using a feature in a decision process may be rooted in prevailing cultural or social norms, or political beliefs or legal (privacy) regulations or historical precedents. Unfortunately, existing outcome-based fairness notions developed for learning systems fail to capture this intuitive human understanding of fairness. Instead, current fair learning mechanisms justify the means (process) by the ends (outcomes), ignoring the different levels of societal consensus on the desirability of using different features in decision making. We propose different notions of process fairness to address this situation.

1.1 Contributions

We highlight the main contributions of this paper.

- We introduce three scalar measures of process fairness.
We operationalize these measures by assembling user preferences for features in the context of (i) recidivism risk estimation using the public COMPAS ProPublica dataset [21], and (ii) prediction of illegal weapon possession using the New York Police stop-question-and-frisk (SQF) dataset [5].

We model the tradeoff between process fairness and accuracy of a classifier as fast, scalable, constrained submodular optimization problems over the set of features, and demonstrate good performance empirically.

On the datasets we considered, our results suggest that high process fairness perhaps surprisingly leads also to high outcome fairness, but that some loss of accuracy is unavoidable.

1.2 Related work

Fairness in machine learning. Prior work on fairness in machine learning focused primarily on achieving high decision-making accuracy while ensuring fair outcomes for all sensitive attribute value groups (e.g., males, females). The notions of fairness considered by those studies, and the specifications of sensitive attributes, are often determined by legal notions of discrimination [1, 7] that focus on correlations between sensitive attributes and decision outcomes.

For example, a set of recent studies [13, 19, 27, 28] have proposed methods to optimize classification accuracy while meeting two different outcomes fairness criteria: (i) similar individuals from different sensitive attribute value groups (e.g., males and females) should receive similar classification outcomes, that is, the classification outcomes should be free of disparate impact, and (ii) similar fraction of individuals from each sensitive attribute value group should be assigned the positive classification outcomes, that is, the classification outcomes should be free of disparate impact. Similarly, some recent studies [17, 26] have focused on optimizing classification accuracy while ensuring that error rates (e.g., false positive rates, false negative rates) for different sensitive attribute value groups should be similar, that is, the classification outcomes should be free of disparate mistreatment.

However, in this paper, we present a broader notion of fairness that goes beyond the current binary legal specifications—which mark all features as sensitive or non-sensitives—by examining the extent to which people feel that each input feature (sensitive or non-sensitive) is unfair to use and how using or ignoring a feature would affect the outcomes.

Fairness in other disciplines. We briefly comment on how our work is inspired by and is related to works in other disciplines such as social sciences, moral philosophy, political sciences, and law. In moral philosophy [8]: a deontological approach considers certain moral truths to be absolute regardless of situation or outcome, which corresponds well with our notion of process fairness. In contrast, a teleological or utilitarian approach focuses on the outcomes, which corresponds well with the notion of outcome fairness.

Prior literature in social, economic, legal, and political sciences distinguishing between direct discrimination and indirect discrimination makes similar observations as we do in this paper. These works point out that the “wrong” of direct (process) discrimination should be distinguished from the “wrong” of indirect (outcome) discrimination [3]. Similarly, we argue that when considering decision-making, fairness of the process is distinct from fairness of the outcome.

2 DEFINING PROCESS FAIRNESS

Suppose a learning method, say a classifier C, has been trained to make decisions using a set of features $\mathcal{F}$. Intuitively, the classifier’s decision process would be considered fair by a user $u$ only if she judges the use of every one of the features in the set $\mathcal{F}$ to be fair. We leverage this intuition to define the process fairness of the set of features $\mathcal{F}$ to be the fraction of all users who consider the use of every one of the features in $\mathcal{F}$ to be fair. For convenience of exposition, we sometimes use the phrase “process fairness of a classifier” interchangeably with the “process fairness of the set of features” used by the classifier.

Our process fairness definition relies critically on users’ judgments about the use of individual features when making decisions. Note that a user’s judgment about a feature may change after they learn how using the feature might affect the decision outcomes. For instance, a user who initially considered a feature unfair for use in predicting recidivism risk might change their mind and deem the feature fair to use after learning that using the feature significantly improves the accuracy of prediction. Similarly, learning that using a feature might increase or decrease disparity in decision outcomes for different demographic groups (e.g., whites vs. non-whites or men vs. women) might make a user change their opinion on the fairness of using that feature in decision making.

To capture the above concepts, we define three measures of process fairness: feature-apriori fairness, feature-accuracy fairness and feature-disparity fairness.

Consider a scenario for making some important decision. Let $\mathcal{U}$ denote the set of all members (‘users’) of society, and $\mathcal{F}$ denote the set of all possible features that might be used in the decision making process.

Feature-apriori fairness. For a given feature $f \in \mathcal{F}$, let $\mathcal{U}_f \subseteq \mathcal{U}$ denote the set of all users who consider the feature $f$ fair to use without a priori knowledge of how its usage affects outcomes. Given a particular set of features $\mathcal{F}$, we define

$$\text{feature-apriori fairness} (\mathcal{F}) := \frac{|\bigcap_{f \in \mathcal{F}} \mathcal{U}_f|}{|\mathcal{U}|}.$$  (1)

Feature-accuracy fairness. Let $\mathcal{U}_f^{\text{Acc}} \subseteq \mathcal{U}$ denote the set of all users who consider the feature $f$ fair to use if it increases the accuracy of a designated classifier. Given a set of features $\mathcal{F} \subseteq \mathcal{F}$, we define

$$\text{feature-acc. fairness} (\mathcal{F}) := \frac{|\bigcap_{f \in \mathcal{F}} \text{Cond} (\mathcal{U}_f, \mathcal{U}_f^{\text{Acc}})|}{|\mathcal{U}|}.$$  (2)

where

$$\text{Cond} (\mathcal{U}_f, \mathcal{U}_f^{\text{Acc}}) = \begin{cases} \mathcal{U}_f^{\text{Acc}}, & \text{if } f \text{ increases accuracy} \\ \mathcal{U}_f, & \text{otherwise}. \end{cases}$$  (3)

Feature-disparity fairness. Let $\mathcal{U}_f^{\text{Disp}} \subseteq \mathcal{U}$ denote the set of all users who consider the feature $f$ fair to use even if it increases a specified measure of disparity (we use disparate mistreatment [26], see Section 6, but other measures such as disparate impact could be used) of a designated classifier. Given a set of features $\mathcal{F} \subseteq \mathcal{F}$,
we define

\[
\text{feature-disp. fairness} (\mathcal{F}) := \frac{\left| \bigcap_{f_i \in \mathcal{F}} \text{Cond}(\mathcal{U}_{f_i}, \mathcal{U}_{f_i}^{\text{Disp}}) \right|}{|\mathcal{U}|},
\]

where

\[
\text{Cond}(\mathcal{U}_{f_i}, \mathcal{U}_{f_i}^{\text{Disp}}) = \begin{cases} 
\mathcal{U}_{f_i}^{\text{Disp}}, & \text{if } f_i \text{ increases disparity} \\
\mathcal{U}_{f_i}, & \text{otherwise.}
\end{cases}
\]

\textbf{Note 1}: Our fairness measures in (1), (2) and (4) constitute set functions over the the ground set \(\mathcal{F}\): they take in a subset of features \(\mathcal{F} \subseteq \mathcal{F}\) and assign it a fairness value in the range \([0,1]\). A value of 0 returned by a set function means the feature set (\(\mathcal{F}\)) is completely unfair according to the corresponding measure, and a value of 1 means that the features is completely fair. Along with our three measures of process fairness, we define three corresponding measures of unfairness, each defined as the respective measure subtracted from 1. For example, feature-apriori unfairness (\(\mathcal{F}\)) = 1 – feature-apriori fairness (\(\mathcal{F}\)).

\textbf{Note 2}: Two of our definitions (2) and (4) apply for a given classifier \(C\). That is, the classifier \(C\) also needs to be specified to determine whether the condition function \(\text{Cond}\) is satisfied. Further, to determine if using a feature \(f_i\) increases accuracy or disparity : (i) in practice, we need to specify a minimum threshold level \(\epsilon\) for the increase; and (ii) we need to decide the base set of features with respect to which \(f_i\) increases the property. For (i), we specified the minimum thresholds for the increase of accuracy and disparity to be equal to 5% of the range of values achieved by the null classifier and the classifier that uses the full set of features. By choosing this range, we hope to capture a significant part of the total range of values, achievable by any of the classifiers that can be constructed using that set of features, even though this does not necessarily hold for disparity. For (ii) we choose the base set to be the empty set, i.e., no features. These choices lead to sensible results on our datasets, though other choices are possible and typically lead qualitatively to similar results.

\textbf{Properties of (un)fairness measures}: We now highlight some properties of our fairness and unfairness measures that we rely on critically in Section 4, when we face the challenging task of efficiently optimizing their tradeoffs against accuracy. The properties of interest to us are the supermodularity and submodularity of the fairness and the unfairness measures, respectively.

Let \(\mathcal{F}\) be a finite set and let \(g\) be a set function \(g : 2^\mathcal{F} \to \mathbb{R}\), where \(2^\mathcal{F}\) denotes the power set of \(\mathcal{F}\). Let \(\mathcal{F}_A\) and \(\mathcal{F}_B\) be sets such that \(\mathcal{F}_A \subseteq \mathcal{F}_B \subseteq \mathcal{F}\), and \(f \in \mathcal{F} \setminus \mathcal{F}_B\).

\textbf{Definition 2.1}. The function \(g\) is supermodular if:

\[
g(\mathcal{F}_A \cup \{f\}) - g(\mathcal{F}_A) \leq g(\mathcal{F}_B \cup \{f\}) - g(\mathcal{F}_B).
\]

Intuitively, a set function is supermodular if it exhibits increasing marginal gains. A function is supermodular iff its negative is submodular.

\textbf{Definition 2.2}. The function \(g\) is non-increasing monotone if:

\[
g(\mathcal{F} \cup \{f_i\}) - g(\mathcal{F}) \leq 0, \forall \mathcal{F} \subseteq \mathcal{F}, f_i \in \mathcal{F} \setminus \mathcal{F}.
\]

\textbf{Proposition 2.3}. All three measures of process fairness (feature-apriori, feature-accuracy and feature-disparity) are monotone non-increasing supermodular set functions with respect to features (equivalently, the respective unfairness measures are monotone non-decreasing submodular functions).

\textbf{Proof}. Let \(g\) be any of the three measures of process fairness. We must show that, for any two \(\mathcal{F}_A, \mathcal{F}_B\) such that \(\mathcal{F}_A \subseteq \mathcal{F}_B \subseteq \mathcal{F}\), then for any \(f \in \mathcal{F} \setminus \mathcal{F}_B\), inequality (6) is satisfied.

To ease notation, let \(\mathcal{F}^\text{f} = \bigcap_{f_i \in \mathcal{F}} \mathcal{U}_{f_i}\). Since \(|\mathcal{U}| > 0\) is a constant, to show (6) holds, it is sufficient to show that:

\[
\left|\mathcal{F}^\text{f}_A\right| - \left|\mathcal{F}^\text{f}_A \cap \{f\}\right| \geq \left|\mathcal{F}^\text{f}_B\right| - \left|\mathcal{F}^\text{f}_B \cap \{f\}\right|.
\]

or, since for any \(X\) and \(v, X = (X \cap \{v\}) \cup (X \setminus \{v\})\), that:

\[
\left|\mathcal{F}^\text{f}_A \setminus \{f\}\right| \geq \left|\mathcal{F}^\text{f}_B \setminus \{f\}\right|.
\]

Since \(\mathcal{F}_A \subseteq \mathcal{F}_B\), it must be the case that \(\mathcal{F}^\text{f}_B \subseteq \mathcal{F}^\text{f}_A\). It follows that \(\mathcal{F}^\text{f}_B \cap \{f\}\subseteq \mathcal{F}^\text{f}_B \cap \{f\}\), and also that \(\mathcal{F}^\text{f}_B \setminus \{f\}\subseteq \mathcal{F}^\text{f}_A \setminus \{f\}\), hence (9) holds and \(g\) is supermodular.

For any set \(\mathcal{F} \subseteq \mathcal{F}, f_i \in \mathcal{F} \setminus \mathcal{F}\), it holds that \(\left|\mathcal{F} \cap \{f_i\}\right| \leq \left|\mathcal{F}\right|\), hence equation (7) holds and we conclude that \(g\) is a monotone non-increasing supermodular set function.

\textbf{3 MEASURING PROCESS FAIRNESS}

In this section, we apply the measures of process fairness defined in Section 2 to decision making tasks for two real-world datasets related to U.S. criminal justice system. Since the measurement of process fairness relies on human judgments, we first describe how we surveyed users to gather these judgments.

\textbf{3.1 Gathering human judgments}

For gathering human judgments, we use the Amazon Mechanical Turk (AMT) platform where users (or workers) can volunteer to perform a wide range of online tasks for pay [9, 23].

For each of the datasets, we describe the relevant prediction task, and then, for each feature, we ask users to respond to 3 questions. For example, for the task of predicting risk of recidivism, for the feature \textit{age} we show the users the following text and questions:

\begin{itemize}
  \item We could use information about a person’s \textit{age} when predicting their risk of recidivism.
  \item \textbf{Q. 1}: Do you believe it is fair or unfair to use this information?
  \item \textbf{Q. 2}: Do you believe it is fair or unfair to use this information, if it \textit{increases the accuracy} of the prediction?
  \item \textbf{Q. 3}: Do you believe it is fair or unfair to use this information, if it makes one group of people (e.g. African American people) more likely to be falsely predicted as having a \textit{higher risk of recidivism} than another group of people (e.g. white people)?
\end{itemize}

We intentionally did not define the term \textit{fair} in our questions as the primary goal of the survey is to gather data about users’ intuitive sense of fairness. As our analysis later in this section shows, users fairness judgements reflect many complex considerations.

For a given dataset, we gather responses to the above questions from 200 different AMT workers (that is, each feature is judged by 200 different workers). The answers to these questions for each
feature will be used to measure feature-apriori, feature-accuracy and feature-disparity measures of fairness.

Since the tasks we consider relate to the U.S. criminal justice system, we only recruited workers who are from the U.S. To ensure the quality of the judgments, we only recruited AMT master workers who have a reputation on the AMT platform for performing their tasks reliably [4]. Additionally, we filtered out the fairness judgments from a small fraction (less than 5%) of users, who provided outlier (anomalous) responses, such as marking a feature as unfair to use apriori to knowing their impact, but marked it as fair when it increases disparity.

3.2 Real-world datasets

We gather the responses from AMT workers for the following two datasets.

ProPublica COMPAS dataset. Our first dataset [21] is related to recidivism risk estimation (i.e., predicting whether a criminal offender would commit another offense within a certain future time). The dataset is made public by ProPublica and it contains information about all criminal defendants who were subject to screening by COMPAS, a commercial recidivism risk assessment tool, in Broward County, Florida, during 2013 and 2014. From this dataset, we gathered 9 features that can be used to construct a classifier for recidivism prediction namely, arrest charge description (e.g., grand theft, possession of drugs), charge degree (misdemeanor or felony), number of prior criminal offenses, number of juvenile felony offenses, juvenile misdemeanor offenses, other juvenile offenses, age of the defendant, gender of the defendant and race of the defendant. A subset of these features have been used in a number of recent studies of racial biases in recidivism risk prediction [10, 14, 20, 26]. Finally, the dataset also contains information about whether the defendant did actually recidivate or not.

NYPD SQF dataset. Our second dataset is related to New York Police Department’s Stop-Question-and-Frisk (NYPD SQF) program [2], where police officers stop and frisk civilians on the suspicion of being involved in a criminal activity. The dataset is publicly available [5] and studied by various prior works in the context of outcome fairness [15]. All of the stops made by police officers, including the accompanying circumstances and reasons for the stop, are recorded and made publicly available [5]. In a recent study, Goel et al. [15] designed a learning task based on appropriate features present in the datasets to predict whether a person is carrying an illegal weapon or not. We used the same prediction task and with a similar set of 30 features as considered by Goel et al.

3.3 Analyzing human judgments of fairness

On each dataset, for each feature, we computed the fraction of AMT workers who judged that feature to be fair under each of the knowledge settings described in questions Q. 1, 2 and 3. The results are shown in Figure 1 and Figure 3.

We observe that for each of our three notions of fairness (Q. 1, 2, 3), the fraction of AMT workers who judged each feature to be fair varied significantly across features, with the ranking of features consistent across the three measures.

ProPublica COMPAS dataset. We observe that the features neatly fall into three subsets, with declining levels of reported fairness. The first subset consists of features which are directly related to the issue at hand, such as the nature of the current charge. Next are distantly related features which provide information about the defendant’s past record as a juvenile. The third set contains features which appear unrelated, such as sex and race. With this perspective, the users’ responses may appear reasonable. In addition, note that the first two (most fair) sets contain volitional features, that is they relate to actions which the defendant chose to take, and hence might reasonably be considered predictive of the defendant’s future actions; whereas the third (most unfair) set comprises features which are physiological and beyond the defendant’s control. The third set is often considered protected by law [1]. However, our results provide a more nuanced, scalar view of the judged fairness of features.

Comparing feature-accuracy fairness (Q. 2) to feature-apriori fairness (Q. 1), as expected, the judged fairness of each feature increases (except for the number of prior offenses, where the difference is barely significant). For example, the fraction of workers who judged the ‘race’ feature to be fair when they know that it increases accuracy doubles (0.42) as compared to when no additional knowledge is provided (0.21). Feature-disparity fairness (measured using Q. 3) is significantly lower than for the first two measures, confirming expectations that a potential for disparity lowers human judgment of fairness.

NYPD SQF dataset. We observe similar trends in NYPD SQF dataset: features that are directly related to the issue at hand, such as suspicion of engaging in a violent crime (‘susp. crime’), are rated as more fair than distantly related features, like acting furtively,
which are in turn rated as more fair than unrelated features, such as sex and race. Similar conclusions as for the COMPAS dataset can be made about volitional and physiological features, as well, even though there are some exceptions, since fitting a relevant description, is a directly related, but a physiological feature.

**Dependence on population demographics.** We also investigate whether the perception of process fairness can change across different demographics within our US survey takers. To this end, we request the AMT workers to provide their demographic information regarding their gender, race, and political leaning and examined the fairness results based on these demographic subgroups. Though some differences were present (for example, almost no non-white person rated race as a feature that is fair to use, and fewer females than males rated juvenile features as fair), results were qualitatively similar for male/female and white/non-white (plots omitted due to space constraints). However, we did observe an interesting difference between the workers who marked their political leaning as “very liberal” and “very conservative”. Responses from these two subgroups are shown in Figure 2. For all three notions of process fairness, conservatives rated all of the features as more fair to use than liberals (for the sake of brevity, the plot shows just feature-apriori and feature-disparity fairness for both subgroups). Additionally, if a feature increased disparity, liberal users decreased their perceived fairness substantially more than conservatives. Thus, it appears that liberals are more sensitive to disparate outcomes, which is consistent with literature in the social sciences suggesting that these different political views may relate to different “moral foundations” [16].

### 4 OPTIMIZING FOR PROCESS FAIRNESS

Thus far, we have used human judgments to quantify process fairness of each of the individual features in the ProPublica COMPAS and NYPD SQF datasets. Here, we begin to examine the process fairness of sets of features and their corresponding classifiers, using the definitions from Section 2. While excluding features deemed highly unfair from a classifier’s inputs will increase its process fairness, it may lead to significantly lower prediction accuracy. We empirically analyze this tradeoff between process fairness and accuracy.

**Classifier used throughout this paper: monotonicity and submodularity.** Throughout this paper, we always use logistic regression with L2-regularization as our classifier.\(^1\) We chose this classifier for two reasons: (i) it has been used frequently in earlier works in this area (e.g., [15]); and (ii) it has attractive properties which will facilitate our approach to scalable optimization (described later in this section). We next explain these attractive properties.

For a given set of features \(S \subseteq \mathcal{F}\) (where \(\mathcal{F}\) represents all the features that are available in the dataset under consideration), let \(\text{acc}(S)\) be the set function given by the accuracy of our chosen classifier trained on the feature set \(S\). Intuitively, as features are added to \(S\), we expect that \(\text{acc}(S)\) will rise but with diminishing returns. That is, we might expect that \(\text{acc}(S)\) is monotonically nondecreasing and submodular. In fact, these properties do not always hold exactly though they are almost true. By leveraging the connection between strong convexity and submodularity, Elenberg et al. [12] showed that \(\text{acc}(S)\) is weakly submodular.

**Accuracy-fairness tradeoff.** For each our datasets, we train a set of classifiers. We would like to train one classifier for each possible subset of the features present in the dataset. For the ProPublica COMPAS dataset, which has 9 features, we train all \(2^9 = 512\) different classifiers. However, our NYPD SQF dataset has 30 features, which would lead to training \(1.1\) billion different classifiers, which is unworkable. Hence we selected a subset of the 16 most informative features using L1 feature selection [25],\(^2\) and trained all \(2^{16} = 65,536\) different possible classifiers.

\(^{1}\) Additionally, for all the experiments performed in this paper, we randomly split the data into 50%/50% train / test folds 5 times and report the average statistics.

\(^{2}\) We trained a logistic regression classifier with L1 regularization on all 30 features and selected the 16 features with highest absolute weights from the weight vector. These 16 features also cover the range of process fairness well.
Figure 4 shows plots of feature-apriori process fairness against accuracy (results for feature-accuracy and feature-disparity fairness are similar, not shown due to space constraints). For both datasets, feature sets that correspond to high fairness (represented by points at the right side of the figures) come at the cost of accuracy, though this effect is clearer for the NYPD SQF dataset. On the ProPublica COMPAS dataset, as fairness rises, it is possible still to achieve good accuracy unless the very highest level of fairness is required.

**Fair feature selection.** We define the problem of fair feature selection as one of finding solutions that lie along the upper envelope of points shown in Figure 4. Specifically, the key challenge lies in selecting a subset of features that either (i) optimize for fairness, given a desired accuracy threshold; or (ii) optimize for accuracy, given a desired fairness threshold. More formally:

i. **Maximizing accuracy under (un)fairness constraints:** Consider a dataset \( D \) consisting of \( N \) records (e.g., a record in NYPD SQF dataset correspond to a person who was stopped by the police) where each record has a corresponding feature set \( \mathcal{F} \). Let \( y \) denote the set of corresponding binary class labels for records in \( D \), i.e., \( y_i \in \{-1, 1\} \) (e.g., in NYPD SQF dataset, \( y_i = 1 \) means the person stopped was found to possess an illegal weapon whereas \( y_i = -1 \) denotes otherwise). Assume that \( \mathcal{D}^S \), where \( S \subseteq \mathcal{F} \), denotes the part of the dataset where for all records, only a subset \( S \) of all features \( \mathcal{F} \) is selected. Given this information, one can formulate the problem of training the most accurate classifier subject to process unfairness constraints as:

\[
\begin{align*}
\text{maximize} & \quad \frac{\text{acc}(S)}{\text{unf}(S)} \\
\text{subject to} & \quad \text{unf}(S) \leq t,
\end{align*}
\]

where \( \text{acc}(S) \) and \( \text{unf}(S) \) are set functions of \( S \subseteq \mathcal{F} \), denoting the accuracy and unfairness of the corresponding classifiers. \( t \) is a desired threshold, specifying the maximum level of unfairness (minimum level of fairness) that is tolerable. For logistic regression (and linear classifiers in general), the accuracy, \( \text{acc}(S) \), of a feature set \( S \), is computed as:

\[
\frac{1}{N} \sum_{i=1}^{N} \text{sign}(\theta^T D_i^S > 0) = y_i,
\]

where “\( \text{sign} \)” represents a dot product and \( \theta \) represents the optimal decision boundary parameters learned through empirical risk minimization. For a logistic regression classifier with L2 regularization, the optimal decision boundary \( \theta^* \) can be found by solving the following maximum likelihood problem: \( \theta^* = \arg\max_{\theta} -\sum_{i=1}^{N} \log p(y_i|D_i^S, \theta) + \lambda \| \theta \|_2 \), where \( p(y_i = 1|D_i^S, \theta) = \frac{1}{1 + e^{-\theta^T D_i^S}} \). Let \( \lambda \in \mathbb{R}^+ \) specifies the regularization strength (in our experiments, we use \( \lambda = 1 \)). The unfairness \( \text{unf}(S) \) is defined according to three different notions (feature-apriori, feature-accuracy and feature disparity) presented in Section 2.

The optimization problem (10) can be solved rapidly provided that accuracy and unfairness are monotone and submodular set functions of \( S \). In this scenario, the above optimization formulation matches the canonical form of submodular cost submodular knapsack (SCSK) problem [18], for which rapid approximate solutions have been proposed. Specifically, the iterated submodular-cost knapsack (ISK) algorithm proposed in [18] offers performance bounds (on \( \text{acc} \) and \( \text{unf} \))

\[
\left[1 - e^{-\frac{K_{\text{unf}}}{1 + (K_{\text{unf}} - 1) (1 - K_{\text{unf}})}}\right],
\]

respectively where \( K_{\text{unf}} = \max\{|S| : \text{unf}(S) \leq t\} \) and \( K_{\text{unf}} \) is the curvature of \( \text{unf} \).

We have already showed in Section 2 that unfairness is a monotone submodular function of \( S \). As discussed near the beginning of this section, Elenberg et al. [12] showed that the accuracy of our classifier (L2-regularized logistic regression) is weakly submodular. Since accuracy is strictly neither monotone nor submodular, while solving (10) using the ISK algorithm [18], the performance bounds provided by Iyer et al [18], need not hold precisely. However, as our empirical results on two different datasets in the next section show, in practice, solutions to (10) yielded by these algorithms are very close to the optimum.

ii. **Minimizing unfairness under accuracy constraints:** If instead one would like to achieve the least unfair (most fair) solution under a given accuracy performance constraint,\(^4\) the tradeoff can alternatively be considered as follows:

\[
\begin{align*}
\text{minimize} & \quad \text{unf}(S) \\
\text{subject to} & \quad \text{acc}(S) \geq t,
\end{align*}
\]

Assuming \( \text{acc}(S) \) and \( \text{unf}(S) \) to be submodular set functions, problem 11 matches the canonical form of submodular cost submodular cover (SCSS) problem [18] and can be solved by using the iterated submodular set cover (ISSC) algorithm proposed in [18] which offers performance bounds on \( \text{unf}(S) \) where \( K_{\text{unf}} \) is the curvature of \( \text{unf} \), \( n \) is the number of features and \( H_{\text{acc}} \) is the approximation factor of the submodular set cover using the function \( \text{acc} \) (details on the bounds can be found in [18]). As pointed out before, due to weak-submodularity and non-monotonicity of \( \text{acc} \), these bounds need not hold, but the empirical results obtained in the next section are very close to the optimum.

5 EVALUATION

Here, we evaluate the effectiveness of applying the constrained submodular optimization methods of [18] to problems (10) and (11). We show that empirically these methods rapidly provide near-optimal tradeoffs between process fairness and accuracy on the ProPublica COMPAS and NYPD SQF datasets.

5.1 Experimental setup

We solve problems (10) and (11) using iterated submodular-cost knapsack (ISK) and iterated submodular set cover (ISSC) methods proposed by Iyer et al. [18]. These methods require training classifiers on various feature subsets of the training data. To ensure (near) submodularity of accuracy, we use logistic regression classifiers with L2 regularization [12].

To examine a broad range of tradeoffs between process fairness and accuracy, we obtain solutions for (10) and (11) using multiple thresholds for unfairness and accuracy respectively. Specifically, for each problem, we use 21 different threshold values covering the full range possible with constant step size. For example, for problem (11), we first train a null classifier (i.e., a classifier that uses no features and always predicts the dominant class). Let the

\[^4\]Examples of cases, where for outcome fairness, one is legally bound to ensure that a decision making process yields the fairest solution under certain performance constraints are discussed at length in [7]. Similar arguments can be made for process fairness.
accuracy of this classifier be $acc_{null}$. Similarly we train a classifier that uses all features. Let the accuracy of this classifier be $acc_{all}$. We use a set of 21 accuracy thresholds for (11) which range from $acc_{null}$ to $acc_{all}$, taking constant step size. Each successive threshold increases the minimum accuracy $t$ required, potentially yielding more unfair solutions.

**Performance comparison.** For the ProPublica COMPAS dataset, for each of the problem variants (10) and (11), we compare the performance of the constrained submodular optimization method with the true optimum achieved by brute force exhaustive enumeration over all the $2^k$ possible classifiers.

For the NYPD SQF dataset, we compute the results of the constrained submodular optimization methods on both (i) all 30 features; and (ii) a subset of the 16 most informative features. We also show the result of exhaustive enumeration to compute the true optimum method over the 16 features.

### 5.2 Results

Results for maximizing accuracy subject to unfairness (10) are shown in Figure 5, and results for minimizing unfairness subject to accuracy (11) are in Figure 6. Before examining details of the results, we make high level observations:

- Our fast methods for constrained submodular optimization work very well empirically, achieving results that are close to optimal.
- This is very encouraging since these methods are highly scalable.
- Results for (10) and (11) are qualitatively similar (compare Figures 5 and 6, noting that the axes are flipped since we use the x-axis consistently for the optimization threshold).

**Maximizing accuracy subject to unfairness constraints (10).** For each notion of process fairness defined in Section 2, Figure 5 shows for each dataset, the accuracy attained (y-axis) for approximate submodular as well as optimal (exhaustive search) methods as one varies the thresholds on unfairness (x-axis).

We examine the accuracy/fairness tradeoff in detail for feature-apriori fairness on the ProPublica COMPAS dataset, see Figure 5(a). Other plots show similar patterns. We observe several stages of the tradeoff as the maximum unfairness threshold is varied. For an unfairness threshold of 0 (a perfectly fair classifier), an empty feature set is selected. This achieves zero unfairness and an accuracy of the null classifier (based on the dominant class). When the threshold is raised slightly to 0.05, accuracy rises sharply. This is due to the addition of highly fair and informative features. Specifically, the feature ‘#prior off.’ (number of prior offenses) was added, raising accuracy by 0.1. This jump up in accuracy can also be seen by considering the right side of Figure 4(a), recognizing we are looking for the highest point possible, as fairness is lowered slightly from 1.

As the unfairness threshold rises, accuracy at first improves slowly as features are added which are highly fair and informative, such as ‘arrest charge description’ and ‘arrest charge degree’. When the threshold reaches 0.55, however, ‘age’ is added (while other fair but less informative features such as ‘arrest charge description’ are discarded), leading to another sharp rise in accuracy. ‘Age’ is a significantly unfair yet highly informative feature. As the threshold rises higher, accuracy plateaus since the additional features that can be accommodated do not add substantial predictive power.

**Minimizing unfairness subject to accuracy constraints (11).** Figure 6 presents the minimum unfairness achieved (y-axis) as one increases the minimum acceptable threshold on accuracy (x-axis). Due to space constraints, we show results only for the NYPD SQF dataset, but the ProPublica COMPAS dataset shows similar tradeoffs. We see patterns very similar to those previously observed in Figure 5.

**Discussion.** It is very encouraging that the fast, scalable, constrained submodular optimization methods yield very similar results to the exhaustively computed optima. We comment here on feasibility of the returned solutions. When we maximized accuracy
subject to an unfairness constraint (10), for all runs, the fast method always returned a feasible solution (satisfying the constraint). However, when we minimized unfairness subject to a minimum accuracy threshold, occasionally the fast method returned solutions which were slightly infeasible, i.e., the accuracy of the returned solution was marginally below the given threshold. For the ProPublica COMPAS dataset, this happened for 2/60 runs; the maximum extent of infeasibility was 0.001 (e.g. if an accuracy threshold of 0.65 was specified, then 0.649 was returned). For the NYPD SQF dataset with 16 features, this happened 5/60 times with maximum infeasibility of 0.002. For the full NYPD SQF dataset with 30 features, this happened 20/60 times with maximum infeasibility of 0.013.

These empirical results suggest that it may be wise in practice to prefer the approach for maximizing accuracy subject to unfairness, particularly for datasets with many features. However, we advise caution since in fact, if both unfairness and accuracy were exactly monotone and submodular, then this approach is not theoretically guaranteed to return a feasible solution (though infeasibility would be bounded), whereas when minimizing unfairness subject to accuracy, it is guaranteed to be feasible [18]. Finally, note that of course, if an infeasible solution is returned in practice, then it is simple just to move the threshold slightly and try again.

We highlight that when maximizing accuracy for fairness, as the maximum unfairness constraint is relaxed, we do not simply see more features being gradually added and never removed. Rather, we sometimes see a significant change in the entire feature set used, in order to optimize the objective. Some features, such as ‘age’ or ‘juvenile offenses’, can have very high predictive power yet very low fairness. This suggests further exploration into suitability of certain features for designing fair yet accurate prediction tasks.

Figure 6: [Minimizing unfairness subject to accuracy constraints] Results for NYPD SQF dataset are shown. ProPublica COMPAS dataset shows very similar trends.

---

### 6 PROCESS VS. OUTCOME FAIRNESS

Thus far, our evaluation has focused on the new process fairness measures that we introduced in this work, ignoring outcome fairness measures that have been used in all prior work in fair learning [11, 13, 19, 22, 24, 26–28]. We now examine empirically the correlations between process fairness and outcome fairness, and their tradeoffs with accuracy.

The measure of outcome fairness we consider here attempts to capture difference in misclassification rates between users of different sensitive attribute groups. Specifically, it adds together the difference in false positive rates (FPR) and the difference in false negative rates (FNR) for whites (w) and non-whites (nw). This

\[
\text{outcome fairness} = |\text{FPR}_w - \text{FPR}_{nw}| + |\text{FNR}_w - \text{FNR}_{nw}|
\]

Outcome fairness values can vary between -2 and 0, with zero corresponding to a very fair classifier and -2 corresponding to very unfair classifier. Other measures of outcome fairness (e.g., disparate impact considered by [27]) could be used but for the types of risk assessment analysis we consider, our definition may be more suitable [6, 26].

To study the tradeoff between process fairness, outcome fairness and accuracy, we train classifiers (optimizing for accuracy) with all possible combinations of features for both ProPublica COMPAS (\(2^9 = 512\) classifiers), and our reduced NYPD SQF (\(2^{16} = 65,536\) datasets), and compute these three statistics (accuracy, process and outcome fairness) for all the classifiers. Figure 7 shows the outcome fairness and accuracy vs. the three notions of process fairness defined in Section 2. Note that further to the right on the x-axis indicates a higher level of process fairness, and further up on the y-axis indicates higher outcome fairness. The color of each point indicates the accuracy of the corresponding classifier. We make the following observations:

- As described earlier, we consider a subset of the 16 most informative features for the NYPD dataset, since considering all 30 features present in the dataset would entail training \(2^{30} = 1.1\) billion classifiers which is computationally unworkable.
High process fairness implies high outcome fairness. All the plots in Figure 7 show a distinctive inverted L-shape. Specifically, for these datasets, once the process fairness is beyond a certain value (about 0.5), outcome unfairness is always close to zero. That is, in these datasets, ensuring high process fairness always leads to high outcome fairness. On the other hand, low process fairness (close to zero) corresponds to almost no constraints on which features may be used, and is compatible with a wide range of outcome fairness. We had not expected this result in advance, and do not claim it will always hold, though it is striking that it holds so clearly on both the datasets we examined. Intuitively, process fairness seems a stronger notion, in that it imposes very restrictive constraints on the classifier. Examining our dataset, we found that the requirement of high process fairness is restricting the choice of features to those which many people feel are fair, and this leads to features with very low correlation (measured as mutual information) to race or other very sensitive features, and thereby indirectly induces high outcome fairness. We intend to explore this observation further in future work over additional datasets.

Tradeoff with accuracy. As just observed, for our datasets, ensuring high process fairness always leads to high outcome fairness. So it is sufficient to optimize only the process accuracy / fairness tradeoff (using mechanisms discussed in Section 4) with at least moderately high process fairness threshold, which will thereby also reap high outcome fairness as a by-product. As one example, regarding the process accuracy / fairness tradeoff presented in Figure 5(a) (top left panel): thresholding the process unfairness is below a threshold of 0.5 (i.e. requiring very high process fairness) on the x-axis, leads to high accuracy of 0.64 (compared to the maximum possible accuracy of 0.68) and a very high outcome fairness value of ~0.18 (compared to the lowest possible value of ~2.0) also.

We further remark that mechanisms proposed in prior fair learning works optimizing outcome fairness and accuracy cannot similarly guarantee high process fairness. Also, while these prior works have shown that one can maintain very high classification accuracy (close to the most accurate classifier) and still achieve a very high value of outcome fairness, the same is not true for process fairness. That is, the mechanisms for our datasets, and we imagine for many others, requiring high process fairness will prevent high classification accuracy, since this requirement will force informative features to be dropped (as explained in Section 5).

7 CONCLUSION
A number of recent works have focussed on the topic of fair learning, i.e., how to learn to make decisions fairly? However, the notions and measures of fairness underlying these works have largely been inspired by and restricted to those explored in anti-discrimination literature. In this work, we look at fairness notions beyond discrimination. Specifically, we introduce three new quantitative measures of fairness that explicitly account for users’ moral sense for whether or not it is fair to use an input feature in the decision making process. We show how we can operationalize these notions by gathering human judgements about using features in the context of two real-world scenarios: recidivism risk estimation and prediction of illegal weapon possession. We show how finding a good tradeoff between process fairness and accuracy of a classifier can be modeled as a fast, scalable, constrained submodular optimization problem over the set of features and demonstrate good empirical performance over real-world datasets. Our results show that when we optimize for high process fairness, we also achieve high outcome fairness as a byproduct, suggesting that our notions of process fairness might subsume the notions of outcome fairness considered in prior works.

REFERENCES