On Fairness in Budget-Constrained Decision Making

Michiel A. Bakker
MIT-IBM Watson AI Lab
MIT Media Lab
Cambridge, MA
bakker@mit.edu

Alejandro Noriega-Campero
MIT-IBM Watson AI Lab
MIT Media Lab
Cambridge, MA
noriega@mit.edu

Duy Patrick Tu
MIT-IBM Watson AI Lab
MIT Media Lab
Cambridge, MA
patrick2@mit.edu

Prasanna Sattigeri
MIT-IBM Watson AI Lab
IBM Research
Yorktown Heights, NY
psattig@us.ibm.com

Kush R. Varshney
MIT-IBM Watson AI Lab
IBM Research
Yorktown Heights, NY
krvarshn@us.ibm.com

Alex 'Sandy' Pentland
MIT-IBM Watson AI Lab
MIT Media Lab
Cambridge, MA
pentland@mit.edu

ABSTRACT

The machine learning community and society at large have become increasingly concerned with discrimination and bias in data-driven decision making systems. This has led to a dramatic increase in academic and popular interest in algorithmic fairness. In this work, we focus on fairness in budget-constrained decision making, where the goal is to acquire information (features) one-by-one for each individual to achieve maximum classification performance in a cost-effective way. We provide a framework for choosing a set of stopping criteria that ensures that a probabilistic classifier achieves a single error parity (e.g. *equal opportunity*) and calibration. Our framework scales efficiently to multiple protected attributes and is not susceptible to intra-group unfairness. Finally, using one synthetic and two public datasets, we confirm the effectiveness of our framework and investigate its limitations.

KEYWORDS

algorithmic fairness, equal opportunity, active feature acquisition, budget-constrained decision making

ACM Reference Format:
Michiel A. Bakker, Alejandro Noriega-Campero, Duy Patrick Tu, Prasanna Sattigeri, Kush R. Varshney, and Alex 'Sandy' Pentland. 2019. On Fairness in Budget-Constrained Decision Making. In *Proceedings of KDD '19: Explainable AI (XAI) for Fairness, Accountability, Trasnsparency (KDD '19: Explainable AI (XAI)).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As machine learning-based decision making has become increasingly ubiquitous—e.g., in criminal justice [16], medical diagnosis [15], human resource management [3], credit [11], and insurance [29]—there is widespread concern over how these systems introduce and perpetuate discrimination and inequality. Consequently, substantial work on defining and achieving fairness in machine learning systems has been published in the last few years.

The vast majority of this research has relied on the assumption that all data is readily available or can be acquired at no or little additional costs. In such a setting, the model bases its decision about an individual always on all features. In practice, however, there are many applications where the acquisition of an additional feature leads to a feature-specific cost [17]. Consider a patient entering a hospital seeking diagnosis. Typically, the doctor starts the diagnosis with only a handful of symptoms. From there, the patient undergoes a progressive inquiry by e.g. measuring vitals or procuring lab tests. At each step, absent sufficient certainty, the inquiry continues. Acquiring all features at once using all possible medical tests is prohibitively expensive and time-consuming, so at each time-step the doctor is tasked with choosing the next feature that most efficiently leads to a more confident diagnosis. This setting, *active feature-value acquisition* (AFA), is becoming increasingly ubiquitous and is relevant in a wide range of contexts, from credit and insurance, to employee recruiting, poverty and disaster mapping, and online advertising [8, 17, 19, 22, 28].

The machine learning community has proposed different frameworks for quantifying fairness in machine learning [10, 16, 30], most of which focus on balancing classification errors across protected population subgroups, towards achieving equal false-positive rates (*predictive equality*), equal false-negative rates (*equal opportunity*), or both (*equal odds*). Here, we focus on satisfying equal opportunity, requiring non-discrimination only within the 'favorable' outcome [10], while also extending these results to satisfying predictive equality. We show that our method can jointly achieve either of these error parity measures and calibration for each subgroup (*test-fairness*), a property commonly required of classifiers in real-world settings [5, 25]. We call an estimator *calibrated* if, when we look at the subset of people who receive any given probability estimate $p \in [0, 1]$, we indeed find a p fraction of them to be positive instances of the classification problem.

To ensure that predictions are fair, "optimal" post-processing methods have been proposed that achieve either 1) equal odds, or 2) parity in one error rate (e.g. equal opportunity) and calibration [10, 25]. These methods rely on randomization to attain fairness:

KDD '19: Explainable AI (XAI), August 4–8, 2019, Anchorage, AK

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

they randomize the predictions for a subset of individuals in the advantaged group, and hence increase error rates for that group. By carefully tuning the share of randomized predictions, one ensures equal error rates across groups. Although these methods are effective, they are also unsettling and several objections to them have been put forth, such as inefficiency, pareto-suboptimality, and intra-group unfairness due to the randomization [5, 10, 20, 25].

Despite the pervasiveness of AFA systems and the recent spike in work on algorithmic fairness, only one paper in the literature has explored fairness at its intersection with AFA [22]. In that work, optimization is used to find an information budget for each population subgroup such that an AFA classifier achieves either 1) one error parity and calibration or 2) equal odds. Notably, by using this additional degree of freedom, they show that one can achieve these notions of fairness in an AFA setting without resorting to randomization.

Our goal is to further investigate the relationship between equalizing error rates and AFA. In particular, we derive a set of stopping criteria that ensures single error parity (equal opportunity or predictive equality) for calibrated probabilistic classifiers. In contrast to previous work, this method does not rely on optimization but directly relates the stopping criteria to the subgroup-specific base rate and the desired error rate. We demonstrate that our framework is effective in practice using one synthetic and two public datasets and show how it extends naturally to a situation with many subgroups defined over multiple protected attributes.

Finally, the method provides an interesting new perspective on two central topics in the fairness literature: *individual fairness* and *fairness gerrymandering*. First, as statistical notions of fairness like equal opportunity are defined with respect to groups, they only provide guarantees to the group average, not to any individual. Individual fairness tries to tackle this issue by using constraints that bind at the individual level [7]. Our method finds a set of stopping criteria that lead to a personalized budget and set of available features for each individual. Hence, intuitively, it trades off inequality (the model and the budgets are personalized and thus different across individuals) for equity (each of the subgroups defined over the set of protected attributes has the same expected false-negative rate and even within subgroups each individual is classified with similar confidence). This can be seen as an attempt to combine statistical and individual notions of fairness as the stopping criteria lead to increased equity at the individual level. Second, in *fairness gerrymandering*, a classifier appears to be fair when measured across each protected attribute but violates the fairness constraint on a subgroup defined over several protected attributes [14]. In contrast to methods based on optimization, our framework is robust against fairness gerrymandering since it ensures all subgroups have the same expected false-negative rate.

2 RELATED WORK

Active feature-value acquisition. Several methods for AFA have been explored ranging from heuristics-based feature acquisition strategies to more recent reinforcement learning methods in which one jointly trains the classifier and the agent that decides which feature to select next [8, 17, 19, 28]. In line with most prior work in AFA, we select the next best feature using a feature acquisition strategy while separately training the classifier. The feature acquisition strategy is based on maximizing the expected utility

$$EU(x_j) = \int_v P(x_j = v)\frac{U(x_j = v)}{c_j} \tag{1}$$

where $P(x_j = v)$ is the probability that feature x_j will take on value v and $U(x_j = v)$ the utility of the model after adding x_j to feature vector $\mathbf{x}$. The utility function could be defined in multiple ways depending on the objective, such as the expected classification error or the expected entropy. We experimented with multiple definitions for utility and found that one that maximizes the expected increase or decrease in probability outputted by the model is most cost-efficient; see Section 5.1 for details.

Fairness in machine learning. Most recent work in fairness in machine learning, including this work, focuses on matching error rates (false-positive or false-negative) across population subgroups. There are, however, multiple other ways to define fairness such as *demographic parity*, *individual fairness*, *fairness through unawareness*, and *counterfactual fairness*. Please refer to [30] for a comprehensive overview of definitions. Methods for achieving fairness fall into three categories [1]. First, there are methods for pre-processing and improving collection of training data [2, 4, 27]. Second, there are methods for constraining the model during training or optimization including methods for fair representation learning [21, 33]. Finally, there are a number of methods for post-processing probabilities to achieve fairness [10, 32]. For achieving equal opportunity and calibration previous post-processing work has relied on randomization which led to an inefficient and pareto-suboptimal classifier [5, 25]. In this work, we post-process a classifier trained on all features by selecting a specific subset of features for each individual.

3 PROBLEM SETUP

The setup of our framework most follows the one in [25] for fairness in the context of calibrated probabilistic classifiers. However, we extend their framework for use in the AFA setting. Let $(\mathbf{x}, y) \sim P$ be an individual in P represented by a d-dimensional feature set and a binary label $y \in \{0, 1\}$. In the AFA setting, $\mathbf{x}^{(q)} \subset \mathbf{x}$ denotes a query on a subset of features in $\mathbf{x}$, with $q \subset \{0, ..., d\}$, and $\mathbf{x}^{(q)}$ the partial feature vector. The decision maker incurs a cost for the collected features $c^{(q)} = \sum_{j \in q} c_j$. The cost vector $\mathbf{c}$ represents the cost of each feature and is the same for each individual in P. It can represent different types of costs that the decision maker or an individual might incur when a feature is queried such as monetary and privacy costs.

We study the context in which a decision maker can choose what information to collect about an individual in order to maximize accuracy while ensuring fairness. Across all individuals in P, the decision maker is constrained by an average information budget:

Definition 1. *The information budget $\bar{b}$ is a global constraint that represents the average budget that can be used across individuals in P, $\bar{b} = \frac{1}{n}\sum_{i \in P} b_i$ with $b_i = \sum_{j \in q} c_j$, the information budget used for feature collection for a single individual i in P.*

In our population P we have a set of k disjoint population subgroups $G_1, \ldots, G_k$ defined over the protected attributes (such as

certain combinations of protected attribute values like race and gender) across which we measure fairness. Note that the number of population subgroups is exponential in the number of protected attributes (e.g. three binary protected attributes will lead to $2^3 = 8$ subgroups). Generally, these subgroups will have different base rates μ_t, which represents the probability of belonging to the positive class $\mu_t = P_{(\mathbf{x},y)\sim G_t}[y = 1]$ across individuals in group t. For classification, we have a separate probabilistic classifier for each group G_t, $h_t : \mathbb{R}^k \to [0, 1]$. In practice, these separate classifiers are stemming from a single classifier trained on P and only differ because of subgroup-specific calibration. For the probabilistic error rates as well as for measuring disparity, we follow the generalized definitions introduced in [25]:

Definition 2. *The generalized false-positive rate for classifier h_t is $c_{fp}(h_t) = \mathbb{E}_{(\mathbf{x},y)\sim G_t}[h_t(\mathbf{x}^{(q)}) \mid y = 0]$. The generalized false-negative rate is $c_{fn}(h_t) = \mathbb{E}_{(\mathbf{x},y)\sim G_t}[1 - h_t(\mathbf{x}^{(q)}) \mid y = 1]$.*

If the classifier would output binary predictions instead of probabilities, these rates would simply represent standard false-positive and false-negative rates. Similarly, we use generalized notions of equalized odds and equal opportunity for probabilistic classifiers:

Definition 3. *Equal opportunity for a set probabilistic classifiers $h_1, \ldots, h_k$ for groups $G_1, \ldots, G_k$ requires $c_{fn}(h_t) = c_{fn}(h_{t'})$ for all possible combinations of t and t'. Equal odds requires both $c_{fn}(h_t) = c_{fn}(h_{t'})$ and $c_{fp}(h_t) = c_{fp}(h_{t'})$.*

For probabilistic classifiers, however, these two conditions do not ensure fairness if the classifier probabilities the classifier outputs are not calibrated. This is confirmed both theoretically and experimentally in [5, 6, 25].

Definition 4. *A classifier h_t is calibrated if $P_{(\mathbf{x},y)\sim G_t}[y = 1 \mid h_t(\mathbf{x}^{(q)}) = p] = p$.*

In Figure 1, we observe the set of calibrated classifiers for two groups G_1 and G_2. For each group, the classifiers lie on a line with slope $(1 - \mu_t)/\mu_t$ that connects the perfect classifier at the origin with the base rate classifier on the $c_{fp} + c_{fn} = 1$ line. The perfect classifier always assigns the correct prediction, while the base rate classifier has no predictive power and naively assigns the base rate to each individual [16, 25]. For an AFA classifier, the base rate classifier is simply the classifier before any features have been acquired $h(\mathbf{x}^{q=\emptyset})$.

4 EQUAL OPPORTUNITY

We will now derive a set of stopping criteria for each population subgroup that ensure satisfying equal opportunity. Intuitively, the stopping criteria should be chosen such that we collect more features for subgroups for which the model is less certain. By stopping later, we acquire more features, have more predictive power, and move down the slope in Figure 1 towards the perfect classifier at the origin. First, we reformulate c_{fn} from Definition 2 as

$$c_{fn}(h_t) = \frac{1}{\sum_{(\mathbf{x},y)\in G_t} \mathbb{1}_{y=1}} \sum_{(\mathbf{x},y)\in G_t} \mathbb{1}_{y=1}(1 - h_t(\mathbf{x}^{(\mathbf{q})})) \quad (2)$$

The normalization can simply be replaced by a constant $1/(|G_t|\mu_t)$ since we marginalize over all $\mathbf{x}$ in G_t. Because we do not have access to ground truth labels $\mathbb{1}_{y=1}$ at test time, we replace them with the estimates from the the probabilistic classifier $h_t(x^{(q)})$:

$$c_{fn}(h_t) = \frac{1}{|G_t|\mu_t} \sum_{(\mathbf{x},y)\in G_t} h_t(\mathbf{x}^{(q)})(1 - h_t(\mathbf{x}^{(\mathbf{q})})). \quad (3)$$

One way to satisfy equal opportunity is to ensure that, in expectation, we have the same generalized false-negative rate c_{fn} for each group G_t such that $\mathbb{E}_{(\mathbf{x},y)\sim G_t}[c_{fn}(\mathbf{x}^{(q)})] = \beta \; \forall t$, where β can be chosen according to the information budget $\bar{b}$. To achieve this, we slowly increase the confidence of our classifier ($h_t(\mathbf{x}^{(q)}) \to 1$ or $h_t(\mathbf{x}^{(q)}) \to 0$) by sequentially adding features one-by-one. We stop collecting features when our probabilistic classifier crosses an upper or lower threshold probability, $h_t(\mathbf{x}^{(q)}) \geq \alpha_u$ or $h_t(\mathbf{x}^{(q)}) \leq \alpha_l$. For a desired β we can find these stopping thresholds α_u and α_l by ensuring equal $h_t(\mathbf{x}^{(q)})(1-h_t(\mathbf{x}^{(\mathbf{q})})/\mu_t = \beta$ for every individual in our group G_t. Bringing everything except the classifier to one side of the equation, we want the probabilities to be $h_t(\mathbf{x}^{(q)}) = \frac{1}{2} \pm \frac{1}{2}\sqrt{1 - 4\beta\mu_t}$ which leads to thresholds

$$\alpha_u = \frac{1}{2} + \frac{1}{2}\sqrt{1 - 4\beta\mu_t}, \qquad \alpha_l = \frac{1}{2} - \frac{1}{2}\sqrt{1 - 4\beta\mu_t} \quad (4)$$

Thus, by choosing the right stopping criteria for each individual $\mathbf{x}$ according to their subgroup-specific base rate μ_t, we ensure that we satisfy equal opportunity. See Figure 1 for an example with two subgroups. In practice, a decision maker would not choose the target rate β but, instead, tune β to meet an information budget $\bar{b}$. A higher information budget $\bar{b}$ allows for a lower target rate β.

Analogously, if we instead want to achieve equalized false-positive rates (predictive equality) across groups, we can derive a similar but different set of thresholds $\alpha_u = \frac{1}{2} + \frac{1}{2}\sqrt{1 + 4\beta(\mu_t - 1)}$ and $\alpha_l = \frac{1}{2} - \frac{1}{2}\sqrt{1 + 4\beta(\mu_t - 1)}$. Finally, to achieve equal odds, we would have to find the same set of thresholds for both equal false-positive rates and equal false-negative rates. The only case for which these thresholds are the same is for $1 - \mu_t = \mu_t$ (i.e. $\mu_t = 0.5$) which is the trivial case for which there was already no unfairness. This confirms the conclusion in [25] that for different base rates, one cannot simultaneously achieve equal odds and calibration.

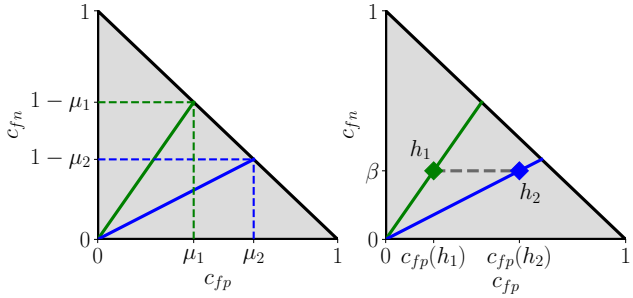


Figure 1: Left, we observe the set of calibrated classifiers h_1 and h_2 for G_1 in green and G_2 in blue. The base rates are $\mu_1 = 0.4$ and $\mu_2 = 0.65$. Right, we observe two classifiers h_1 and h_2 that satisfy calibration and equal opportunity with a target generalized false-negative rate β.

Table 1: Overview of the datasets and subgroups split by the protected attributes. Accuracy and AUC are computed on a dataset-level using the full feature set, while μ is the dataset-level base-rate $P(y)$. For each subgroup we compute the relative number of individuals n_t and the base rate μ_t.

Dataset						Subgroup_1			Subgroup_0		
Name	$N_{samples}$	N_{feat}	Acc	AUC	μ	Label_1	n_1	μ_1	Label_0	n_0	μ_0
Synthetic [9, 23]	10,000	150	85.9%	0.933	50.0%	$z = 1$	50.0%	41.3%	$z = 0$	50.0%	58.9%
Mexican poverty [12, 22]	70,305	182	78.7%	0.856	35.5%	Urban	63.6%	34.9%	Rural	36.4%	36.6%
Adult income [18]	49,000	14	86.3%	0.911	23.9%	White	85.4%	25.4%	Non-white	14.6%	15.3%

Assumptions. In this framework, we make two key assumptions. First, we assume that for each individual we have sufficient statistical power to reach the target β by simply adding more features. In practice, however, there will be a non-zero Bayes-optimal error rate such that we cannot reach the perfect classifier with $\beta = 0$ even with unlimited budget for feature acquisition. Second, we assume that the probabilities are exactly $p = \alpha_u$ or $p = \alpha_l$ while in reality we stop when we cross the threshold and thus $p \geq \alpha_u$ or $p \leq \alpha_l$. In the experiments in Section 5 we show that relaxing both assumptions does not limit the effectiveness of our framework.

4.1 Implications

This result is important for several reasons. First, it provides a theoretical framework for understanding the results presented in [22]. In the *active fairness* framework described there, optimization is used to find a set of parameters that allows for equal opportunity and calibration in the AFA setting, but lacks a theoretical underpinning.

Second, we only need a subgroup's base rate to find α_u and α_l. This is crucial when the problem is extended to a case with several multi-class protected attributes, like gender, race, and sexual orientation. If one instead would try to find the parameters by optimizing over a budget and fairness constraints for each protected attribute, the resulting classifier could contain intra-group unfairness.

Third, comparing this result to the randomization approach presented in [25], our framework shows that by using AFA, we can achieve fairness in a budget-constrained setting without having to resort to randomized approaches that are inefficient, pareto-suboptimal, and lead to intra-group unfairness.

5 EXPERIMENTS

In light of these findings, we demonstrate the effectiveness and limitations of our framework on one synthetic and two public real-world datasets. In this section we aim to satisfy equal opportunity (equal false-negative rates) but in Appendix A we demonstrate that the method can also be used for satisfying predictive equality (equal false-positive rates).

5.1 Implementation

Implementation requires two elements, a probabilistic model and a feature acquisition strategy.

Probabilistic model. First, we need a model that allows us to estimate $P(y|\mathbf{x}^{(q)})$ for arbitrary feature subsets $\mathbf{x}^{(q)}$, with $q \in [0, d]$. We implement this using a probabilistic random forest, designed to deal with incomplete data in trees [26]. Specifically, we first train a standard random forest using the complete feature vector $\mathbf{x}$ for each individual in our training set. At test time, however, we now only have access to part of the feature vector $\mathbf{x}^{(q)}$. In a probabilistic random forest, when the algorithm encounters a tree node for which the value is missing in the feature vector $\mathbf{x}^{(q)}$, the algorithm continues along both branches towards the leafs while the outcomes in each branch are weighted based on the estimated probability for the missing value. For each individual, that probability is estimated from the frequency of values in the training set. We then compute classification probabilities as a weighted average of the leaf purity across all leaves landed on by the search. Finally, the predicted probability is averaged across all trees. Analogously, gradient boosting and other models can be adjusted to admit incomplete feature vectors [26, 31]. In this work, all random forests are created using `scikit-learn` with 64 trees and maximally 150 leaf nodes. Additionally, we built a custom predict function that works with the `scikit-learn` object but accounts for the missing feature values.

Feature acquisition strategy. Second, we implement an efficient feature acquisition strategy to estimate which next feature can be best selected based on the current partially observed feature vector $\mathbf{x}^{(q)}$, while balancing cost and increasing accuracy. We implement a *greedy* feature selection algorithm based on the expected utility methods described in [13, 17]. For an individual with feature vector $\mathbf{x}$, and at each feature collection iteration, the algorithm searches for the feature $j' \notin q$ that maximizes the difference between the current predicted probability $\hat{P}$ and the expected probability given that an additional feature j' is queried with cost c_j, given by:

$$j' = \underset{\{j: j \notin q,\, j \in [0,d]\}}{\arg\max} \frac{1}{c_j} \left| \hat{P}\{y = 1|\mathbf{x}^{(q \cup j)}\} - \hat{P}\{y = 1|\mathbf{x}^{(q)}\} \right|. \quad (5)$$

5.2 Datasets

An overview of the datasets is given in Table 1. All results are computed using random 60%/20%/20% train/validation/test splits. The Synthetic dataset is generated using the `make_classification` function from `scikit-learn` [9, 23] where we use the default set of parameters while setting `class_sep` to 1.5 (default is 1.0) to make the task slightly easier. The protected attribute is a randomly selected feature which we exclude from the dataset and binarize by splitting along the median. The Mexican Poverty dataset is extracted from the 2016 publicly available Mexican household survey containing household binary poverty levels for prediction, as well as a series of household features [12]. We will release the processed

dataset. Finally, we use the Adult Income dataset from UCI Machine Learning Repository [18] which comprises demographic and occupational attributes, with the goal of classifying whether a person's income is above $50,000.

5.3 Achieving equal opportunity

We empirically demonstrate that our framework satisfies equal opportunity for a given information budget $\bar{b}$. To make the results more interpretable, we choose the costs to be the same for each feature $c_j = 1$. Hence, the budget $\bar{b}$ is simply the average number of features that can be queried across individuals. We also tested the framework with linearly increasing feature costs and feature costs drawn from a normal distribution, while observing similar behavior. To ensure calibrated probabilities, we fit a sigmoid function to the classifier's scores using the validation set; a calibration method known as Platt scaling [24].

Figure 2 demonstrates that we can satisfy equal opportunity for the three datasets. In Figure 2a we plot the derived equal opportunity classifiers in the generalized false-positive/false-negative plane with 5%, 10%, and 20% as information budgets for respectively the Synthetic, Adult Income, and Mexican Poverty datasets. Table 2 shows the residual false-negative disparities after applying our stopping criteria as well as the false-positive disparity. The results in the table are benchmarked against the disparities between groups when the classifiers have access to all features (i.e. no stopping criteria). As expected, our framework leads to drastically lower false-negative disparities while the false-positive disparities are similar to the baseline. In Table A1 we find the stopping criteria, budgets, and classifier performance (using area under the ROC curve) for equal opportunity, predictive equality and the baseline model using all features.

Figure 2b demonstrates that our framework allows for achieving equal opportunity for a range of different information budgets. The steepest decrease in c_{fn} is observed for smaller information budgets because our feature acquisition strategy chooses the most predictive features first. For larger budgets, the curves plateau as the additional features do not further increases the classifier performance.

Our framework assumes a one-to-one mapping between the target β and the actual generalized-false negative rate c_{fn}. For the Synthetic dataset in Figure 2c, we indeed observe an approximate one-to-one mapping between target and actual. For the Mexican Poverty dataset, however, we observe a strong positive correlation but the actual false-negative rate increases slower than the target rate. For small target rates β this is the result of lower overall classification performance for the Mexican dataset (AUC 0.85 when using all features versus 0.93 for the Synthetic dataset); as β becomes smaller, the thresholds α_u and α_l approach 1 and 0. Therefore, when the classification performance is low, many instances will fail to meet the stopping criteria before running out of possible features to query which increases the actual rate c_{fn}. For high values of β, another effect is at play. The smaller than expected c_{fn} is observed because our probabilities do not end up exactly at α_u and α_l (our stopping criteria are defined as $h_t(\mathbf{x}^{(\mathbf{q})}) \geq \alpha_u$, not as $h_t(\mathbf{x}^{(\mathbf{q})}) = \alpha_u$). Importantly, however, we observe that these assumptions do not affect our ability to achieve equal opportunity.

Table 2: Information budgets $\bar{b}$ and absolute differences (disparities) in generalized false-negative $|\Delta c_{fn}|$ and false-positive rates $|\Delta c_{fp}|$ for the equal opportunity classifiers visualized in Figure 2. We benchmark our framework to the classifiers with access to all features x ($\bar{b} = 100\%$).

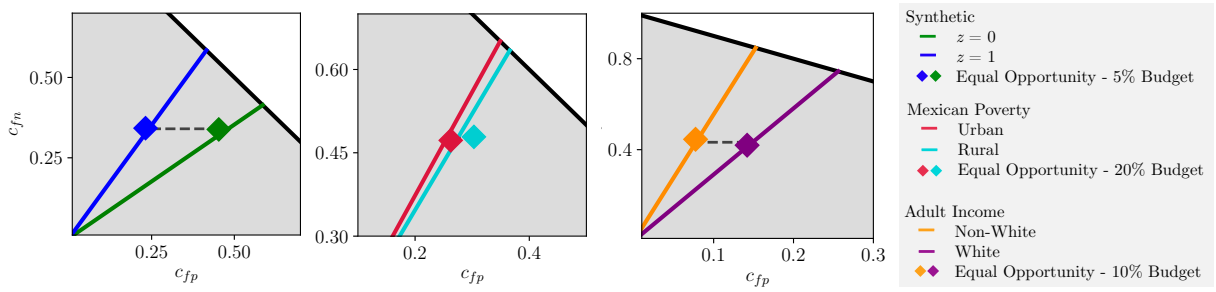
	Equal opportunity			All features		
Dataset	$\bar{b}$	$\lvert\Delta c_{fn}\rvert$	$\lvert\Delta c_{fp}\rvert$	$\bar{b}$	$\lvert\Delta c_{fn}\rvert$	$\lvert\Delta c_{fp}\rvert$
Synthetic	5%	0.0039	0.221	100%	0.097	0.042
Mexican Pov.	20%	0.0063	0.040	100%	0.019	0.042
Adult income	10%	0.026	0.065	100%	0.038	0.056

Finally, we test our framework for eight disjoint subgroups defined over three protected household attributes (Young/Old, Urban/Rural, and With/Without Children) in the Mexican Poverty dataset. When using optimization for achieving fairness, large intra-group unfairness can manifest itself; even though disparities measured across protected attributes are small, large differences between false-negative rates for each subgroup defined over the attributes can exist [14]. In contrast, our framework requires all eight false-negative rates to be approximate equal and, indeed, empirically we observe that all fall within the $[0.440, 0.541]$ range. See Tables A2 and A3 in the appendix for an overview of results for the three protected attributes.

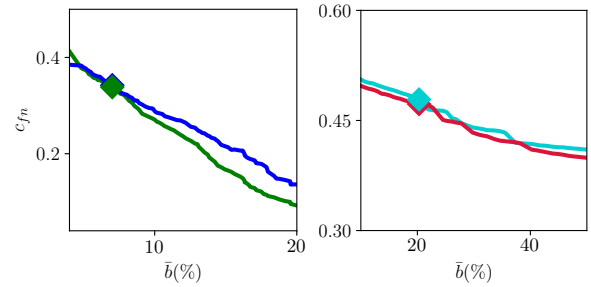
6 CONCLUSION

We introduced a framework for achieving equal opportunity (and predictive equality) for calibrated probabilistic classifiers in an active feature-value acquisition setting. The framework relates a target generalized false-negative rate and a subgroup-specific base rate to a set of stopping criteria, used to determine when to stop querying additional features for fair classification. The target false-negative rate can be tuned using the available information budget. The relationship between error and base rates is intuitive as base rate differences are what give rise to disparities between calibrated classifiers. On three datasets, we show the effectiveness of the framework and demonstrate that relaxing some of the assumptions in our framework does not significantly change its effectiveness.

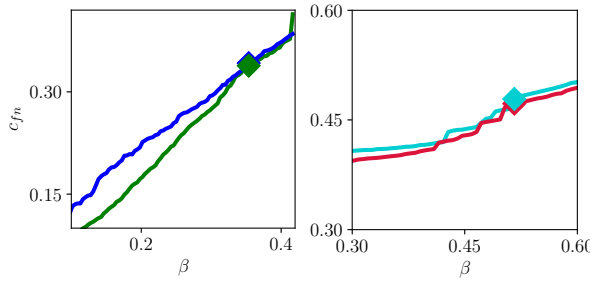
Importantly, the proposed framework neither relies on optimization nor any form of randomization. Furthermore, it is not susceptible to intra-group unfairness and provides a new perspective on how we could combine individual and statistical notions of fairness. The ability to set the expected false-negative rates for each subgroup simply by deriving a set of stopping criteria could be used to ensure statistical notions of fairness to hold not only for a small number of larger subgroups but potentially for an exponential number of smaller subgroups. This could enable a set of classifiers for which both individual and statistical notions of fairness hold without having to collect the protected attributes. In turn, this allows for fair decision making in contexts where one deals with a multitude of subgroups or when collecting the protected attributes is unethical or impossible.



(a) The line of calibrated classifiers and the equal opportunity classifiers plotted in the generalized false-positive/false-negative plane similar to Figure 1. The values for the differences between the error rates can be found in Table A3. The black line traces $c_{fp} + c_{fn} = 1$ and contains the naive base rate classifiers for which no features are queried.



(b) Generalized false-negative rates c_{fn} for equal opportunity classifiers along a range of different information budgets $\bar{b}$.



(c) Generalized false-negative rates c_{fn} for different target false-negative rates β. Ideally, you expect a straight-line relationship with slope 1.

Figure 2: For the datasets described in Table 1, we demonstrate equal opportunity for three different budgets. For each subgroup, we show the possible set of calibrated classifiers (lines) together with the specific classifier that achieves equal opportunity for the given budget (diamonds).

REFERENCES

[1] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2019. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv* (2019).

[2] Flavio P. Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems*. 3992–4001.

[3] Aaron Chalfin, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. Productivity and selection of human capital with machine learning. *American Economic Review* 106, 5 (2016).

[4] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why Is My Classifier Discriminatory?. In *Advances in Neural Information Processing Systems*. 3539.

[5] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv* (2018).

[6] Cynthia S Crowson, Elizabeth J Atkinson, and Terry M Therneau. 2016. Assessing calibration of prognostic risk scores. *Statistical methods in medical research* 25, 4 (2016), 1692–1706.

[7] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 214–226.

[8] Tianshi Gao and Daphne Koller. 2011. Active classification based on value of classifier. In *Advances in Neural Information Processing Systems*.

[9] Isabelle Guyon. 2003. Design of experiments of the NIPS 2003 variable selection benchmark.

[10] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315.

[11] Cheng-Lung Huang, Mu-Chen Chen, and Chieh-Jen Wang. 2007. Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications* 33, 4 (2007), 847–856.

[12] Pablo Ibarrarán, Nadin Medellín, Ferdinando Regalia, Marco Stampini, Sandro Parodi, Luis Tejerina, Pedro Cueva, and Madiery Vásquez. 2017. How Conditional Cash Transfers Work. (2017).

[13] Pallika Kanani and Prem Melville. 2008. Prediction-time active feature-value acquisition for cost-effective customer targeting. *Advances In Neural Information Processing Systems (NIPS)* (2008).

[14] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2017. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv* (2017).

[15] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction policy problems. *American Economic Review* 105, 5 (2015).

[16] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint* (2016).

[17] Balaji Krishnapuram, Shipeng Yu, and R Bharat Rao. 2011. *Cost-sensitive Machine Learning*. CRC Press.

[18] Moshe Lichman et al. 2013. UCI machine learning repository.

[19] Li-Ping Liu, Yang Yu, Yuan Jiang, and Zhi-Hua Zhou. 2008. TEFE: A time-efficient approach to feature extraction. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE.

[20] Pranay K. Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R. Varshney, and Ruchir Puri. 2019. Bias Mitigation Post-Processing for Individual and Group Fairness. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2847–2851.

[21] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. 2015. The variational fair autoencoder. *arXiv* (2015).

[22] Alejandro Noriega-Campero, Michiel Bakker, Bernardo Garcia-Bulle, and Alex Pentland. 2019. Active Fairness in Algorithmic Decision Making. *Proceedings of AAAI / ACM Conference on Artificial Intelligence, Ethics, and Society* (2019).

[23] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.

[24] John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*

10, 3 (1999), 61–74.
[25] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*. 5680–5689.
[26] Maytal Saar-Tsechansky and Foster Provost. 2007. Handling missing values when applying classification models. *Journal of machine learning research* 8, Jul (2007), 1623–1657.
[27] Prasanna Sattigeri, Samuel C. Hoffman, Vijil Chenthamarakshan, and Kush R. Varshney. 2019. Fairness GAN: Generating Datasets with Fairness Properties using a Generative Adversarial Network. In *ICLR Workshop on Safe Machine Learning*.
[28] Hajin Shim, Sung Ju Hwang, and Eunho Yang. 2018. Joint active feature acquisition and classification with variable-size set encoding. In *Advances in Neural Information Processing Systems*. 1368–1378.
[29] Eric Siegel. 2013. *Predictive analytics: The power to predict who will click, buy, lie, or die*. Wiley Hoboken.
[30] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.
[31] David Williams, Xuejun Liao, Ya Xue, and Lawrence Carin. 2005. Incomplete-data classification using logistic regression. In *Proceedings of the 22nd International Conference on Machine learning*. 972–979.
[32] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. 2017. Learning non-discriminatory predictors. *arXiv* (2017).
[33] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. *arXiv preprint* (2017).

A ACHIEVING PREDICTIVE EQUALITY

In line with satisfying equal opportunity in the main text, we empirically demonstrate that our framework satisfies predictive equality (equal false-positive rates) for three different information budgets 10%, 15%, and 30% for respectively the Synthetic, Adult Income, and Mexican Poverty datasets. In Table A1 we observe the statistics for both equal opportunity and predictive equality. In agreement with equal opportunity, we see a drastic decrease in target error rate (now $|\Delta c_{fp}|$) with respect to the false-positive disparity measured across the benchmark classifiers that have access to all features.

Table A1: Comparison of AUC, absolute differences in generalized false-negative $|\Delta c_{fn}|$ and false-positive $|\Delta c_{fp}|$ rates across the equal opportunity, predictive equality and benchmark classifiers for the three different datasets. The equal opportunity and predictive equality classifier were derived by setting a group-specific threshold and applying active feature acquisition while the benchmark classifier has access to the complete feature set. The upper threshold α_u is shown while the lower threshold relates to the upper threshold as $\alpha_l = 1 - \alpha_u$. Both are determined by the average information budget $\bar{b}$.

	Equal opportunity						Predictive equality						All features			
Dataset	$\bar{b}$	$\lvert\Delta c_{fn}\rvert$	$\lvert\Delta c_{fp}\rvert$	AUC	$\alpha_{u,1}$	$\alpha_{u,0}$	$\bar{b}$	$\lvert\Delta c_{fn}\rvert$	$\lvert\Delta c_{fp}\rvert$	AUC	$\alpha_{u,1}$	$\alpha_{u,0}$	$\bar{b}$	$\lvert\Delta c_{fn}\rvert$	$\lvert\Delta c_{fp}\rvert$	AUC
Synthetic	5%	0.0039	0.221	0.77	0.82	0.71	10%	0.225	0.002	0.77	0.69	0.81	100%	0.097	0.042	0.933
Mexican Poverty	20%	0.0063	0.040	0.78	0.77	0.75	30%	0.038	0.011	0.79	078	0.79	100%	0.019	0.042	0.856
Adult Income	10%	0.026	0.065	0.86	0.78	0.89	15%	0.423	0.010	0.81	0.78	0.73	100%	0.038	0.056	0.911

Table A2: Active feature acquisition for eight different subgroups defined over three binary protected attributes in the Mexican Poverty dataset. The metrics c_{fn}, c_{fp} and AUC are computed on each subgroup level with a 25% information budget $\bar{b}$. Each subgroup has its own threshold as stopping criterion based on the subgroup specific base rate μ_t. Furthermore, we report the relative number of individuals n_t with respect to the whole set and the fairness statistics for the benchmark case.

								Equal opportunity			All features		
Subgroup						n_t	μ_t	c_{fn}	c_{fp}	AUC	c_{fn}	c_{fp}	AUC
Young	$\cap$	Urban	$\cap$	With Children		20.0 %	51.4%	0.465	0.460	0.667	0.305	0.306	0.848
Young	$\cap$	Urban	$\cap$	Without Children		13.4%	21.6%	0.502	0.174	0.828	0.494	0.140	0.866
Young	$\cap$	Rural	$\cap$	With Children		13.5%	50.3%	0.443	0.446	0.699	0.333	0.349	0.812
Young	$\cap$	Rural	$\cap$	Without Children		5.8%	23.0%	0.541	0.186	0.773	0.559	0.153	0.810
Old	$\cap$	Urban	$\cap$	With Children		7.4%	54.0%	0.448	0.468	0.681	0.320	0.307	0.750
Old	$\cap$	Urban	$\cap$	Without Children		22.4%	21.7%	0.543	0.188	0.810	0.542	0.160	0.838
Old	$\cap$	Rural	$\cap$	With Children		4.5%	49.5%	0.440	0.433	0.711	0.339	0.322	0.817
Old	$\cap$	Rural	$\cap$	Without Children		12.7%	24.1%	0.530	0.224	0.785	0.531	0.191	0.804
$\cup_{Subgroups}$						100%	35.5%	0.480	0.283	0.794	0.432	0.233	0.824

Table A3: Absolute differences in generalized false-negative $|\Delta c_{fn}|$ and false-positive $|\Delta c_{fp}|$ rates on a group-level. The thresholds for feature acquisition were set on a subgroup level (same as in Table 4). Controlling for error rates (in this case c_{fn}) on a subgroup level leads to fairness on the level of the sensitve attribute.

Group	$\lvert\Delta c_{fn}\rvert$	$\lvert\Delta c_{fp}\rvert$
Young/Old	0.045	0.088
Urban/Rural	0.070	0.097
With/Without Children	0.089	0.257