Fair Enough: Improving Fairness in Budget-Constrained Decision Making Using Confidence Thresholds

Michiel Bakker,[1,2] **Humberto Riverón Valdés,**[1,2] **Duy Patrick Tu,**[1,2] **Krishna P. Gummadi,**[3]
Kush R. Varshney,[4,2] **Adrian Weller,**[5,6] **Alex 'Sandy' Pentland**[1,2]
[1]Massachusetts Institute of Technology, Cambridge, USA
[2]MIT-IBM Watson AI Lab, Cambridge, USA
[3]Max Planck Institute for Software Systems, Saarbrücken, Germany
[4]IBM Research, Yorktown Heights, USA
[5]University of Cambridge, Cambridge, UK
[6]The Alan Turing Institute, London, UK

Abstract

Increasing concern about discrimination and bias in data-driven decision making systems has led to a growth in academic and popular interest in algorithmic fairness. Prior work on fairness in machine learning has focused primarily on the setting in which all the information (features) needed to make a confident decision about an individual is readily available. In practice, however, many applications allow further information to be acquired at a feature-specific cost. For example, when diagnosing a patient, the doctor starts with only a handful of symptoms but progressively improves the diagnosis by acquiring additional information before making a final decision. We show that we can achieve fairness by leveraging a natural affordance of this setting: the decision on when to stop acquiring more features and proceeding to predict. First, we show that by setting a single set of confidence thresholds for stopping, we can attain equal error rates across arbitrary groups. Second, we extend the framework to a set of group-specific confidence thresholds which ensure that a classifier achieves equal opportunity (equal false-positive or false-negative rates). The confidence thresholds naturally achieve fairness by redistributing the budget across individuals. This leads to statistical fairness across groups but also addresses the limitation that current statistical fairness methods fail to provide any guarantees to individuals. Finally, using two public datasets, we confirm the effectiveness of our methods empirically and investigate the limitations.

Introduction

Recent work on fairness in machine learning-based decision making has focused on predictive models that make decisions when all data is readily available or can be acquired at little additional cost. In such a setting, the model makes a classification decision for each individual based on all features. In practice, however, there are many scenarios in which the acquisition of an additional feature leads to a feature-specific cost for the decision maker (Krishnapuram, Yu, and Rao 2011). 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, so at each time-step, the doctor is tasked with acquiring the next piece of information that most efficiently leads to a confident diagnosis. This setting, known as *prediction-time active feature-value acquisition* (AFA), is relevant in a wide range of contexts, from credit assessment, to employee recruiting, poverty and disaster mapping, and advertising (Gao and Koller 2011; Liu et al. 2008; Shim, Hwang, and Yang 2018; Krishnapuram, Yu, and Rao 2011).

At the same time, the machine learning community has proposed myriad definitions for fairness (Verma and Rubin 2018), that can be broadly categorized in two groups. (1) *Statistical* definitions of fairness focus on balancing classification errors across protected population subgroups, towards achieving equal error rates (*overall accuracy equality*), equal false-positive rates (*predictive equality*), equal false-negative rates (*equal opportunity*), or both (*equal odds*). Although these notions are simple and can be easily verified, they fail to give any meaningful guarantees to individuals or subgroups within the protected groups. (2) *Individual* notions of fairness, on the other hand, formalize constraints that bind on the individual-level, as opposed to a quantity that is averaged over a group. For example, (Dwork et al. 2012) requires that 'similar individuals should be treated similarly'. Unfortunately, the need for a good task-specific distance metric has prevented this definition from being used in practice. In this work, we demonstrate that by using confidence thresholds in AFA, we can give fairness guarantees at both the group and the individual level.

Despite the pervasiveness of AFA systems and the recent interest in algorithmic fairness, to our knowledge only one study has explored fairness at its intersection with AFA (Noriega-Campero et al. 2019). In that work, an optimization method is used to find an information budget for each population subgroup such that an AFA classifier achieves parity in false-positive or false-negative rates. Notably, by using the information budget as an additional degree of freedom during optimization, they show that several statistical notions of fairness can be achieved in an AFA setting. Our work is different in that it provides a novel framework for mitigating both group and individual unfairness using confidence thresholds. In particular, we derive a set of stopping criteria for AFA which ensure that we only classify an indi-

vidual's outcome once we have acquired a sufficient number of features to a certain level of confidence. Because the level of confidence will be the same across individuals, we attain error parity for calibrated probabilistic classifiers across groups and in expectation also across individuals.

While our method for achieving fairness is different from earlier work, we suggest that it is more intuitive in many settings as it trades off inequality (the set of features that are used for decision making are personalized and thus different across individuals) for equity (each of the individuals are classified with equal confidence). When a decision maker encounters an individual from a subgroup that it has less experience with, e.g. because the group is underrepresented in the training set, more information needs to be collected to make a fair decision with a similar level of confidence.

Moreover, our method has interesting implications on the privacy of the individuals. While it may appear unreasonable to require more information from those in underrepresented groups, in fact — in contrast to methods that necessitate all features to be collected before making a prediction — our algorithm only acquires the smallest possible set of features to reach a desirable level of confidence. Therefore it naturally follows the 'data minimization' principle as expressed in Article 5(1)(c) of the EU's General Data Protection Regulation (GDPR) which provides that personal data shall be 'adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed'.[1]

Our main contribution is the formulation of confidence thresholds, which we provide for attaining equal error rates, equal false-positive rates and equal false-negative. Experimentally, we demonstrate that our framework is effective using two public datasets.

Related work

Prediction-time active feature-value acquisition

An AFA system consists of three components: 1) a classifier that can handle partially observed feature sets, 2) a strategy for determining which feature to select next based on the features that are already collected, and 3) a stopping criterion for determining when to stop acquiring more features and make a final prediction. First, there are different ways that classifiers handle partial features sets. For discriminative models, feature imputation is a model-agnostic way of handling missing data but there exist also more effective model-specific methods. For the tabular datasets we consider in this work, we found the best performance using distribution-based imputation for random forests in which the possible assignments of missing values are weighted proportionally (Saar-Tsechansky and Provost 2007).

Second, to determine which feature to select next, we need a method that estimates the cost-effectiveness of each of the unselected features based on the features we have already selected. For simplicity and in line with most prior work on AFA, we use a heuristic method that maximizes the expected utility of a feature, where the utility function is based on the expected increase in the absolute difference between the estimated class probabilities of the two most likely classes (Kanani and Melville 2008). However, our framework is acquisition method-agnostic and works with any strategy. A more recent approach, Efficient Dynamic Discovery of High-Value Information (EDDI), uses a partial variational autoencoder to represent the partial feature set of already acquired features. It then computes the mutual information between the current representation and each of the available features to select the feature that minimizes this information (Ma et al. 2019).

Finally, to determine when to stop selecting additional features, most prior work assumes some given feature budget per individual such that the decision maker is tasked with selecting the most cost-effective features within that budget (Krishnapuram, Yu, and Rao 2011). The *active fairness* framework extends this to group-specific budgets to attain statistical notions of fairness (Noriega-Campero et al. 2019).

Fairness in machine learning

Most recent work on fairness in machine learning focuses on *statistical fairness* by matching error rates (false-positive, false-negative or accuracy) across protected subgroups. *Overall accuracy equality* is achieved when the total classification error is the same across protected subgroups (Berk et al. 2018). The measure is only useful when true positives and negatives are equally desirable but it is nonetheless studied in the scientific literature and in the ProPublica analysis of COMPAS (Chen, Johansson, and Sontag 2018; Larson et al. 2016). Second, one can consider equal false positive rates (*predictive equality*) and false negative rates (*equal opportunity*) when either one of them is desirable (Hardt et al. 2016). We refer to (Verma and Rubin 2018) for an overview of definitions.

In contrast, *individual fairness* definitions have no notion of protected subgroups, but instead formulate constraints that bind on pairs of individuals (Dwork et al. 2012; Joseph et al. 2016). Both families of definitions have strength and weaknesses. Statistical notions do not provide any guarantees to individuals, while individual notions have obstacles to deployment and require strong assumptions on agreed-upon fairness metrics.

Two recent papers, (Kearns et al. 2017) and (Hébert-Johnson et al. 2018), attempt to combine the 'best of both worlds' by asking for statistical definitions to hold on an exponential class of groups defined by a class of functions of bounded complexity. Although promising, the approach has proven to be difficult to implement and ultimately still inherits the weaknesses of statistical fairness at a smaller scale (Chouldechova et al. 2018).

Problem setup

Let $(\mathbf{x}^{(i)}, y^{(i)}) \sim P$ be individual i in P represented by a d-dimensional feature set and a binary label $y^{(i)} \in \{0, 1\}$. In the AFA setting, we acquire the features in sequential order starting with the empty set $\mathcal{O}_0 := \emptyset$ at time $t = 0$. At every later timestep t we choose a subset of features from the unselected set of features, $\mathbf{S}_t^{(i)} \subseteq \{1, \ldots, d\} \setminus \mathcal{O}_{t-1}^{(i)}$ and examine the value of $\mathcal{S}_t^{(i)}$ at a cost $c_t^{(i)} := \sum_{j \in \mathcal{S}_t^{(i)}} c_j$. After

[1] https://eur-lex.europa.eu/eli/reg/2016/679/2016-05-04

each new acquisition step, the classifier will have access to features $\mathcal{O}_t^{(i)} := \mathcal{S}_t^{(i)} \cup \mathcal{O}_{t-1}^{(i)}$. We keep acquiring features up to time $T^{(i)}$ when we meet a stopping criterion. At that point, we will classify $\mathbf{x}^{(i)}$ using only the set of features in $\mathcal{O}_{T^{(i)}}^{(i)}$. Note that the specific set of selected features $\mathcal{O}_{T^{(i)}}^{(i)}$ will be highly dependent on the individual i. The cost vector $\mathbf{c}$ is equal for every individual in P and can represent different types of costs such as monetary or privacy costs.

A decision maker acquires a unique set of features $\mathcal{O}_{T^{(i)}}^{(i)}$ for every individual i where the specific feature set is optimized in order to balance the expected quality of the final decision with the total costs of the features.

Fairness

In our population P, let us assume we have a set of disjoint subgroups G_a with $a \in \mathcal{A}$, which, for example, could represent subgroups split by race or gender. Generally, these subgroups can have different base rates μ_a, which represents the probability of belonging to the positive class $\mu_a = P[y = 1 \mid A = a]$. For classification, we train a separate probabilistic classifier for each group G_a, $h_a : \mathbb{R}^k \to [0, 1]$, predicting the probability that the individual has binary label $y = 1$. In practice, these separate classifiers are stemming from a single classifier trained on P and only differ because of subgroup-specific calibration. The classifiers allow for classification of partial feature sets $h_a(\{x_j\}_{j \in \mathcal{O}_t})$ which we write as $h_a(\mathcal{O}_t)$ for brevity. For the probabilistic error rates as well as for measuring disparity, we follow the generalized definitions introduced in (Pleiss et al. 2017):

Definition 1. *The generalized false-positive rate for classifier h_a is $c_{fp}(h_a) = \mathbb{E}_{(\mathbf{x},y)\sim G_a}[h_a(\mathcal{O}_T) \mid y = 0]$. The generalized false-negative rate is $c_{fn}(h_a) = \mathbb{E}_{(\mathbf{x},y)\sim G_a}[1 - h_a(O_T) \mid y = 1]$. The generalized error rate is equivalent to the L_1 loss $c_{err}(h_a) = \mathbb{E}_{(\mathbf{x},y)\sim G_a}[|y - h_a(O_T)|]$*

If the classifier would output binary predictions $h \in \{0, 1\}$ instead of probabilities, these rates would simply represent standard false-positive rates, false-negative rates, and the zero-one loss. Similarly, we use generalized notions of equal accuracy, equal opportunity, and predictive equality for probabilistic classifiers:

Definition 2. *Equal accuracy for a set of probabilistic classifiers h_1 and h_2 for groups G_1 and G_2 requires $c_{err}(h_1) = c_{err}(h_2)$. Similarly, predictive equality requires $c_{fp}(h_1) = c_{fp}(h_1)$ and equal opportunity $c_{fn}(h_1) = c_{fn}(h_2)$.*

Exact equality is hard to enforce in practice so we study the degree to which these constraints are violated: $|c_{fp,1} - c_{fp,2}|, |c_{fn,1} - c_{fn,2}|, |c_{err,1} - c_{err,2}|$. Furthermore, for probabilistic classifiers, these fairness conditions only hold if the classifier probabilities are calibrated. This is confirmed both theoretically and experimentally in (Pleiss et al. 2017).

Definition 3. *A classifier h_a is calibrated if $P_{(\mathbf{x},y)\sim G_a}[y = 1 \mid h_a(\mathcal{O}_t) = p] = p$.*

In Figure 1, we observe the set of calibrated classifiers for two groups G_1 and G_2. For each group, the set of calibrated classifiers $h \in \mathcal{H}$ 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 (Pleiss et al. 2017). 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. For an AFA classifier, the base rate classifier represents the classifier before any features have been acquired $h_a(\emptyset) = \mu_a$.

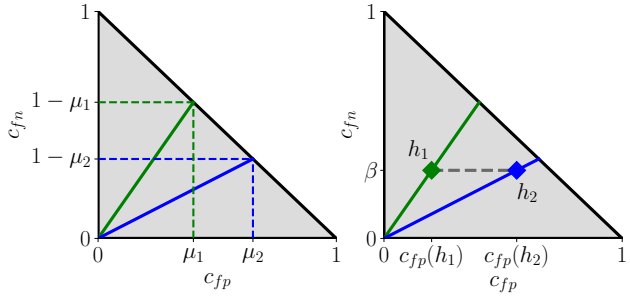

Figure 1: Left, we observe the set of calibrated probabilistic classifiers h_1 and h_2 for G_1 in green and G_2 in blue. The base rates are $\mu_1 = 0.4$ and $\mu_2 = 0.65$. Right, we observe two classifiers h_1 and h_2 that satisfy and equal opportunity with a target generalized false-negative rate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

Confidence thresholds

Intuitively, the stopping criteria should be chosen such that we collect more features for individuals and groups 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. For different measures of fairness, we will derive an upper and lower confidence threshold α_u and α_l. The upper threshold corresponds to predicting $y = 1$ with confidence α_u while the lower threshold corresponds to predicting $y = 0$ with confidence $1 - \alpha_l$. We reach these thresholds by sequentially adding features one-by-one, slowly increasing the confidence of our classifier ($h_a(\mathcal{O}_t) \to 1$ or $h_a(\mathcal{O}_t) \to 0$). We stop collecting features when the probability meets either one of the thresholds, $h_a(\mathcal{O}_t) \geq \alpha_u$ or $h_a(\mathcal{O}_T) \leq \alpha_l$.

In the framework that follows, we make three key assumptions. First, we assume that for each individual we have sufficient relevant features to reach any threshold by simply adding more features. In most real-world datasets, there will be a non-zero classification error even when all features are collected such that, for some individuals, we will not reach thresholds close to 0 or 1 even with unlimited budget for feature acquisition. To address this issue, decision makers can either choose thresholds closer to the base rate or leverage the model's ability to select a unique set of features for each individual and collect more features that are relevant for the set of individuals that are currently hard to classify. Second, we assume the probabilities after stopping to be exactly $p = \alpha_u$ or $p = \alpha_l$ while in reality we stop when we cross the threshold and thus find $p \geq \alpha_u$ or $p \leq \alpha_l$. When

this overshooting effect is stronger for one of the groups, this could lead to unfairness. Finally, throughout this work, we will treat the calibration constraint as holding exactly. In the Supplementary Material (SM), we also present the confidence thresholds for approximately calibrated classifiers. Despite relying on these assumptions, we show in the Experiments section that our framework mitigates disparity in real-world datasets.

Equal error rates

We will derive a set of stopping criteria for each subgroup that ensure satisfying equal error rates (similar to *overall accuracy equality* in previous work (Verma and Rubin 2018)). We first rewrite the expected c_{err} from Definition 1. We write $\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}$ and $\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}$ as $\mathbb{E}_{\mathrm{G_a}}$ and $\mathrm{P}_{\mathrm{G_a}}$ when it is clear from the context.

$$\begin{aligned} c_{err}\left(h_a\right) &= \mathbb{E}_{\mathrm{G_a}}\left[\left|\, h_a(\mathcal{O}_\mathrm{T}) - y\right|\right] \\ &= \int_0^1 p\mathrm{P}_{\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T}) = p \mid y = 0\right] \mathrm{P}_{\mathrm{G_a}}\left[y = 0\right] + \\ &\quad (1-p)\,\mathrm{P}_{\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T}) = p \mid y = 1\right] \mathrm{P}_{\mathrm{G_a}}\left[y = 1\right] dp \end{aligned}$$

Now we apply Bayes rule to find

$$\begin{aligned} c_{err}\left(h_a\right) = \int_0^1 \big(&p\mathrm{P}_{\mathrm{G_a}}\left[y = 0 \,|\, h_a(\mathcal{O}_\mathrm{T}) = p\right] + (1-p) \\ &\mathrm{P}_{\mathrm{G_a}}\left[y = 1 \,|\, h_a(\mathcal{O}_\mathrm{T}) = p\right]\big)\mathrm{P}_{\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T}) = p\right] dp \end{aligned}$$

Substiuting $\mathrm{P}_{\mathrm{G_a}}\left[y = 0 \mid h_a(\mathcal{O}_\mathrm{T}) = p\right] = 1 - p$ and $\mathrm{P}_{\mathrm{G_a}}\left[y = 1 \mid h_a(\mathcal{O}_\mathrm{T}) = p\right] = p$ results in[2]

$$\begin{aligned} c_{err}\left(h_a\right) &= \int_0^1 2\left(p^2 - p\right) \mathrm{P}_{\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T}) = p\right] dp \\ &= 2(\mathbb{E}_{\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})\right] - \mathbb{E}_{\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})^2\right]) \end{aligned}$$

To attain equal error rates, we want to ensure equal $c_{err}(h_a(\mathcal{O}_T))$ in expectation for all individuals, i.e., $c_{err}(h_a(\mathcal{O}_T^{(i)})) = \beta_{err}, \forall a \in \mathcal{A}$. For a desired β_{err} we can find the stopping thresholds α_u and α_l by ensuring equal $2(\mathbb{E}_{\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})\right] - \mathbb{E}_{\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})^2\right]) = \beta_{err}$ for every individual in group G_a.

$$\begin{aligned} \beta_{err} &= 2(\mathbb{E}_{\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})\right] - \mathbb{E}_{\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})^2\right]) \\ &= 2p_u(\alpha_u - \alpha_u^2) + 2p_l(\alpha_l - \alpha_l^2) \end{aligned}$$

where p_u and p_l are the probabilities that we reach the upper or lower thresholds averaged over G_a and $p_u + p_l = 1$. When $\alpha_l + \alpha_u = 1$, the solution for the confidence thresholds follows as

$$\alpha_u = \frac{1}{2} + \frac{1}{2}\sqrt{1 - 2\beta_{err}} \quad \alpha_l = \frac{1}{2} - \frac{1}{2}\sqrt{1 - 2\beta_{err}} \quad (1)$$

If, for every individual, we acquire features one-by-one until we reach either of these thresholds, we achieve, in expectation, an equal error for every individual. Importantly, these thresholds are independent of the subgroup label a and will therefore lead to equal error rates for any subgroup $a \in \mathcal{A}$ as long as the probabilities are calibrated with respect to subgroup a.

[2]Here, we assume perfect calibration. The derivation for approximate calibration can be found in the SM.

Equal false-positive or false-negative rates

When the desired measure of fairness is equal false-positive rates (predictive equality) or equal false-negative rates (equal opportunity), the thresholds derived for equal error rates will not suffice as each group has a different base rate μ_a. To derive a new set of thresholds we first reformulate c_{fp} from Definition 1 and follow a derivation similar to that for equal error rates.

$$\begin{aligned} c_{fp}\left(h_a\right) &= \mathbb{E}_{\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T}) \mid y = 0\right] \\ &= \int_0^1 p\ \mathrm{P}_{\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T}) = p \mid y = 0\right] dp \\ &= \int_0^1 p\ \frac{1 - \mathrm{P}_{\mathrm{G_a}}\left[y = 1 \,|\, h_a(\mathcal{O}_\mathrm{T}) = p\right]}{1 - \mathrm{P}_{\mathrm{G_a}}\left[y = 1\right]} \mathrm{P}_{\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T}) = p\right] dp \end{aligned}$$

Using $\mathrm{P}_{\mathrm{G_a}}\left[y = 1 \mid h_a(\mathcal{O}_\mathrm{T}) = p\right] = p$ and $\mathrm{P}_{\mathrm{G_a}}\left[y = 1\right] = \mu_a$, we can rewrite this as

$$\begin{aligned} c_{fp}\left(h_a\right) &= \frac{1}{1-\mu_q}\int_0^1 p(1-p)\,\mathrm{P}_{\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T}) = p\right] dp \\ &= \frac{1}{1-\mu_a}(\mathbb{E}_{\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})\right] - \mathbb{E}_{\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})^2\right]) \end{aligned}$$

Following the same steps for the false-negative rate, we find

$$c_{fn}\left(h_a\right) = \frac{1}{\mu_a}(\mathbb{E}_{\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})\right] - \mathbb{E}_{\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})^2\right])$$

We define a target false positive rate β_{fpr} to find the stopping criteria for each group such that $c_{fp}(\mathcal{O}_T) = \beta_{fpr}$ for all groups G_a. We then find a set of stopping criteria, analogously to those for c_{err},

$$\alpha_u = \frac{1}{2} + \frac{1}{2}\sqrt{1 - 4\beta_{fpr}(1-\mu_a)} \quad (2)$$

$$\alpha_l = \frac{1}{2} - \frac{1}{2}\sqrt{1 - 4\beta_{fpr}(1-\mu_a)} \quad (3)$$

For false-negative rates we find a similar but different set of stopping criteria for a target false-negative rate β_{fnr}

$$\alpha_u = \frac{1}{2} + \frac{1}{2}\sqrt{1 - 4\beta_{fnr}\mu_a} \quad (4)$$

$$\alpha_l = \frac{1}{2} - \frac{1}{2}\sqrt{1 - 4\beta_{fnr}\mu_a} \quad (5)$$

Note that all individuals within a group will stop at the same thresholds. Our method therefore prevents intra-group unfairness, an often cited limitation of other statistical fairness methods, ignoring fair assignment of outcomes within a sensitive subgroup (Grgić-Hlača et al. 2017; Kearns et al. 2017). Moreover, the thresholds now differ per group, and we can therefore not achieve fairness with respect to arbitrary unlabeled subgroups with different base rates. However, when unfairness is caused only by a difference in variance across groups, for example because of different sample sizes $|G_a|$ or because of group-conditional feature variance $\mathrm{Var}(\mathbf{x}|a)$, the base rates across group will be equal, leading to fairness across even unknown subgroups.

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
Mexican poverty	70,305	182	78.7%	0.856	35.5%	Urban	63.6%	34.9%	Rural	36.4%	36.6%
Adult income	49,000	14	86.3%	0.911	23.9%	White	85.4%	25.4%	Non-white	14.6%	15.3%

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_a and the base rate μ_a.

Experiments

We demonstrate the effectiveness and limitations of our framework on two public real-world datasets. In this section we aim to minimize the generalized error and generalized FPR disparity while results for FNR can be found in the SM. In each experiment we select different information budgets by testing different values of the target error rates β_{err}, β_{fp}, and β_{fn}. In turn, this changes the upper and lower confidence thresholds α_u and α_l. When the thresholds α_l and α_u are set closer to respectively 0 and 1, the feature acquisition stops later leading to a higher average budget use. By using confidence thresholds, more budget will automatically be allocated to subgroups for which the classifier generally has less confidence, which mitigates disparity. We benchmark the results in each experiment against equally distributing the budgets across groups which we call 'equal budget'. For the different target error rates in the experiments that follow, β_{err}, β_{fn} and β_{fp}, we first calculate the average budget consumed when using the 'confidence thresholds' method and then distribute that budget equally across all individuals to obtain the benchmark. In this way, we are able to benchmark the disparity for different information budgets that a decision maker could have available. We measure the overall performance of the classifiers using the Area Under the Receiver Operating Characteristics curve (AUC) to account for the imbalanced label distributions.

Implementation

In addition to the confidence-based stopping criteria, implementation requires two more elements: a probabilistic model and a feature acquisition strategy. First, we need a model that allows us to estimate $P(y|\mathcal{O}_t)$ for arbitrary feature subsets. Although implementing this is easier with generative models like Naive Bayes, we use distribution-based imputation in random forest (Saar-Tsechansky and Provost 2007) as random forest has superior predictive performance (AUC 0.83 using the full feature on the Mexican Poverty dataset set versus 0.79 for Naive Bayes). Specifically, we first train a standard random forest using the complete feature vectors $\mathbf{x}$. At prediction time, when the algorithm encounters a tree node for which the value is missing in the feature set $\mathcal{O}_t$, it continues along both branches towards the leafs while the outcomes in each branch are weighted based on the estimated probability for the missing value. We then compute the probabilities using a weighted average of the leaf purity across all leaves landed on by the search. Finally, the probability is averaged across all trees. All random forests are created using `scikit-learn` with 64 trees and maximally 150 leaf nodes. Additionally, we built a custom function that accounts for the missing feature values, which we will make public.

Second, we implement a feature acquisition strategy to estimate which next feature should be selected based on the current feature set $\mathcal{O}_t$, while balancing cost and increasing accuracy. We implement a *greedy* feature selection algorithm based on the expected utility method introduced in (Kanani and Melville 2008). For an individual with observed feature set $\mathcal{O}_t$, and at each iteration of the feature collection process, the algorithm searches for the feature $j' \notin \mathcal{O}_t$ that maximizes the difference between the current predicted probability P and the expected probability given that an additional feature j' is queried with cost c_j, given by:

$$j' = \arg\max_{j:j\notin\mathcal{O}_t} \frac{1}{c_j} \sum_v P(x_j = v \mid \mathcal{O}_t) \\ |P(y = 1 \mid \mathcal{O}_t \cup \{x_j = v\}) - P(y = 1 \mid \mathcal{O}_t)| \quad (6)$$

where $P(x_j = v| \mathcal{O}_t)$ is estimated from the training dataset.

Finally, a decision maker will generally not reason in terms of a target confidence but instead will have a budget it can spend on average for each individual, $\bar{b} = \frac{1}{n}\sum_{i\in P}\sum_{j\in\mathcal{O}^{(i)}_{T^{(i)}}} c_j$. The cost for each feature c_j can be different and can represent for example monetary or privacy costs. To make the results more interpretable, we choose the costs to be the same for each feature $c_j = 1$. Hence, the budget $\bar{b}$ will simply be the average number of features that can be collected across individuals. Changing these costs to make them more realistic will only lead to a different ordering of features and will not further impact the results in this section. Assuming there will not be a distributional shift between training and test time, we calibrate this average budget by varying the confidence thresholds and observing the budget spent on a hold out set.

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 Mexican Poverty dataset is extracted from a 2016 publicly available Mexican household survey containing household binary poverty levels for prediction, as well as a series of household features (Ibarrarán et al. 2017). We will release the processed dataset. Finally, we use the Adult Income dataset from UCI Machine Learning Repository (Lichman and others 2013) which comprises

demographic and occupational attributes, with the goal of classifying whether a person's income is above $50,000.

Achieving equal error rates

We empirically demonstrate that our framework mitigates the error disparity for the Mexican Poverty dataset in Figure 2 along a range of information budgets. The results for the Adult Income dataset can be found in Figure SM1 in the SM. To ensure calibrated probabilities, we fit a sigmoid function to the classifier's probabilities using a validation set; a calibration method known as Platt scaling (Platt 1999). Crucially, we calibrate across the entire population, effectively ignoring the underlying groups to show that we can mitigate unfairness without explicitly accounting for these subgroups.

The leftmost panel in Figure 2 shows the effectiveness but also the limitations of our framework. For the full range of information budgets our method outperforms the benchmark. For smaller information budgets, we see that the effect is the strongest; there are sufficient relevant features for every individual to reach the thresholds and we thus see that our method strongly mitigates the disparity, despite higher error rates for each group at smaller budgets. As the information budget grows, there are an increasing number of individuals for which the algorithm exhausts all relevant features before we reach the confidence thresholds which limits the effectiveness of our framework. Eventually, when we acquire all features for $\bar{b} = 100\%$ the disparity naturally approaches the disparity for the benchmark. The center panel shows how our framework mitigates disparity by redistributing budget from the the Urban subgroup ($a = 1$) to the Rural subgroup ($a = 0$). Finally, in the rightmost panel, we observe that the performance-disparity trade-off of our method Pareto dominates the benchmark.

Achieving equal false-positive or false-negative rates

Next, we show that our framework mitigates the generalized FPR disparity for the Mexican Poverty dataset in Figure 3 along a range of information budgets. The results for the FNR disparity in this dataset, as well as FPR disparity for the Adult Income dataset can be found in Figure SM2 and Figure SM3 the SMM. We now have access to the sensitive attribute and thus calibrate the probabilities for each group separately, effectively creating separate classifiers for each group.

In the leftmost panel of Figure 3, we find that the disparity-budget trade-off of the confidence thresholds method Pareto dominates the trade-off for the equal budget benchmark. However, for very small and large budgets, we see that the effectiveness is limited. Initially, both thresholds will be close to 0.5 while the classifier for each group starts at the base rate when no features have been collected leading to an immediate stop. Once the threshold crosses the base rate, feature acquisition starts but as the most predictive features will be acquired first this leads to an overshooting of the probabilities, violating the assumption that the probabilities stop exactly at the intended thresholds. As the budget increases further this effect is mitigated and the effectiveness increases. For large budgets, we see the same effect as observed previously when mitigating the error disparity; the algorithm exhausts all relevant features before the thresholds can be reached. Eventually the disparity approaches the benchmark when for both methods all features are collected. In the center panel, we observe how the method mitigates the disparity by redistributing budget from the Urban group to the Rural group. In the rightmost panel, we observe that our method Pareto dominates the equal budget benchmark along the full AUC-disparity trade-off.

Conclusion and Discussion

We introduced a framework for achieving equal error rates, equal opportunity and predictive equality in an active feature-value acquisition setting. The framework relates a target generalized error, false-negative or false-positive rate to a set of confidence thresholds, used to determine when to stop querying features for each individual.

In addition to achieving statistical fairness, our approach can be interpreted as staking a novel middle-ground between individual and statistical fairness. This is most obvious in the case where we have one set of confidence thresholds that effectively leads to equal expected error rates for each individual and hence to equal overall error rates across an arbitrary set of underlying subgroups. However, even when we aim to equalize false positive or false negative rates across groups, and thus use different thresholds for each group, we naturally acquire more information for those individuals for which the classifier faces most uncertainty, leading to equal expected error rates for every member of a protected subgroup. Hence, our framework mitigates intra-group unfairness or 'fairness gerrymandering' that is generally seen as a strong limitation of previous statistical fairness methods (Kearns et al. 2017).

On two public datasets, we show that our method minimizes disparities. Especially for small budgets, our framework strongly mitigates disparities while for large budgets we exhaust the relevant features before reaching the confidence thresholds. This issue also represents a limitation of the datasets we use in this work. The features in both datasets have been carefully chosen to be cost-effective for the majority of individuals in the dataset. In our framework, however, it is natural to add features that are relevant only to a handful of individuals, as they will only be selected for that group. Hence, we encourage future work that investigates the applications of our framework to datasets and settings that meet this criterion and work that extends our method to supporting models that facilitate partial feature sets also during training time.

Finally, we encourage further research that investigates the implications on the privacy of individuals both at training and at prediction time. Generally, we found that active feature acquisition is a natural framework to achieve 'data minimization'; it collects only the minimum set of features. However, even though our method reduces error disparities, it can actually create privacy disparities as for each individual a different set of features will be collected. To address

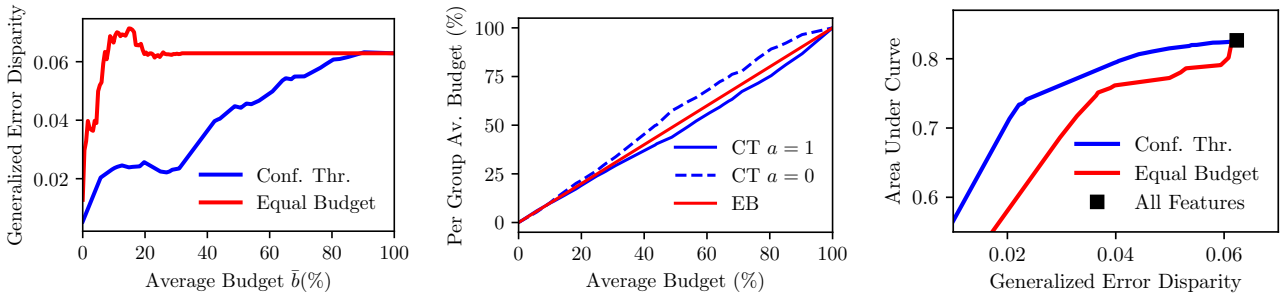

Figure 2: Confidence thresholds for achieving equal generalized error rates across the Urban and Rural subgroups in the Mexican Poverty dataset. In each plot, the curves are generated by sweeping the target error rate β_{err}, changing the average budget allocated to each groups. Left, the residual error disparity for the confidence thresholds (blue) and the equal budget benchmark (red). Center, the average budget per group versus the total average budget. The dashed and solid blue lines represent the average budget used for respectively the Urban ($a = 1$) and Rural ($a = 0$) subgroups while the red line represents the average budget for both groups using the benchmark method. Right: The Pareto front of the AUC versus disparity trade-off for our method and the benchmark method as well as for the classifier with access to all features in black.

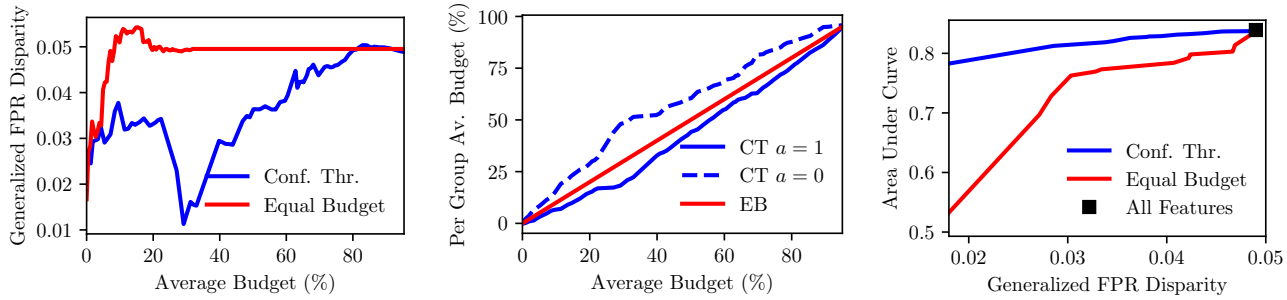

Figure 3: Confidence thresholds for achieving equal generalized FPR rates across the Urban and Rural subgroups in the Mexican Poverty dataset. In each plot, the curves are generated by sweeping the target error rate β_{fpr}. Left, the residual generalized FPR disparity for the confidence thresholds (blue) and the equal budget benchmark (red). Center, the average budget per group versus the total average budget. The dashed and solid blue lines represent the average budget used for respectively the Urban ($a = 1$) and Rural ($a = 0$) subgroups while the red line represents the average budget for both groups when using the benchmark method. Right, the Pareto front of the AUC-disparity trade-off for our method and the benchmark method. The black square represents the classifier that has access to all features.

this, a natural extension would be to work towards a framework that holistically trades-off monetary costs for decision makers, privacy costs for decision subjects, and fairness.

References

Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; and Roth, A. 2018. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 0049124118782533.

Chen, I.; Johansson, F. D.; and Sontag, D. 2018. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems*, 3539.

Chouldechova, A.; Benavides-Prado, D.; Fialko, O.; and Vaithianathan, R. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, 134–148.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226. ACM.

Gao, T., and Koller, D. 2011. Active classification based on value of classifier. In *Advances in Neural Information Processing Systems*.

Grgić-Hlača, N.; Zafar, M. B.; Gummadi, K.; and Weller, A. 2017. On fairness, diversity, and randomness in algorithmic decision making. In *4th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Hardt, M.; Price, E.; Srebro, N.; et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 3315.

Hébert-Johnson, Ú.; Kim, M.; Reingold, O.; and Rothblum, G. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International*

Conference on Machine Learning, 1944–1953.

Ibarrarán, P.; Medellín, N.; Regalia, F.; Stampini, M.; Parodi, S.; Tejerina, L.; Cueva, P.; and Vásquez, M. 2017. *How Conditional Cash Transfers Work*. Number 8159 in IDB Publications (Books). Inter-American Development Bank.

Joseph, M.; Kearns, M.; Morgenstern, J. H.; and Roth, A. 2016. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, 325–333.

Kanani, P., and Melville, P. 2008. Prediction-time active feature-value acquisition for cost-effective customer targeting. *Advances In Neural Information Processing Systems (NIPS)*.

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2017. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv*.

Krishnapuram, B.; Yu, S.; and Rao, R. B. 2011. *Cost-sensitive Machine Learning*. CRC Press.

Larson, J.; Mattu, S.; Kirchner, L.; and Angwin, J. 2016. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)* 9.

Lichman, M., et al. 2013. Uci machine learning repository.

Liu, L.-P.; Yu, Y.; Jiang, Y.; and Zhou, Z.-H. 2008. Tefe: A time-efficient approach to feature extraction. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE.

Ma, C.; Tschiatschek, S.; Palla, K.; Hernandez-Lobato, J. M.; Nowozin, S.; and Zhang, C. 2019. Eddi: Efficient dynamic discovery of high-value information with partial vae. In *International Conference on Machine Learning*, 4234–4243.

Noriega-Campero, A.; Bakker, M.; Garcia-Bulle, B.; and Pentland, A. 2019. Active fairness in algorithmic decision making. *Proceedings of AAAI / ACM Conference on Artificial Intelligence, Ethics, and Society*.

Platt, J. C. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10(3):61–74.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*, 5680–5689.

Saar-Tsechansky, M., and Provost, F. 2007. Handling missing values when applying classification models. *Journal of machine learning research* 8(Jul):1623–1657.

Shim, H.; Hwang, S. J.; and Yang, E. 2018. Joint active feature acquisition and classification with variable-size set encoding. In *Advances in Neural Information Processing Systems*, 1368–1378.

Verma, S., and Rubin, J. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, 1–7. IEEE.

Fair Enough: Improving Fairness in Budget-Constrained Decision Making Using Confidence Thresholds

Supplementary Material

Michiel Bakker,[1,2] **Humberto Riverón Valdés,**[1,2] **Duy Patrick Tu,**[1,2] **Krishna P. Gummadi,**[3]
Kush R. Varshney,[4,2] **Adrian Weller,**[5,6] **Alex 'Sandy' Pentland**[1,2]
[1]Massachusetts Institute of Technology, Cambridge, USA
[2]MIT-IBM Watson AI Lab, Cambridge, USA
[3]Max Planck Institute for Software Systems, Saarbrücken, Germany
[4]IBM Research, Yorktown Heights, USA
[5]University of Cambridge, Cambridge, UK
[6]The Alan Turing Institute, London, UK

Confidence thresholds for approximately calibrated classifiers

This section presents the derivation of the confidence thresholds for approximately calibrated classifiers. Parts of this derivation are adopted from (Pleiss et al. 2017). First, we define approximate calibration.

Definition 4. *A classifier h_a is approximately calibrated with respect to a group G_a if*

$$\int_0^1 \Big| P_{(\mathbf{x},y)\sim G_a}\big[y{=}1 \mid h_a(\mathcal{O}_\mathrm{T}){=}p\big] - p \Big| \, P_{(\mathbf{x},y)\sim G_a}\big[h(\mathbf{x}) = p\big]\, dp \leq \delta_{cal} \tag{7}$$

where $\mathcal{O}_t$ is the feature set at time t and δ_{cal} is the bound on the calibration error. A classifier is perfectly calibrated when $\delta_{cal} = 0$.

Lemma 1. *If for a group G_a the calibration error is bounded by δ_{cal} then*

$$2\left(\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})\big] - \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})^2\big] - \delta_{cal}\right) \leq c_{err}\left(h_a\right) \leq 2\left(\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})\big] - \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})^2\big] + \delta_{cal}\right) \tag{8}$$

$$\frac{1}{1-\mu_a}\left(\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})\big] - \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})^2\big] - \delta_{cal}\right) \leq c_{fp}\left(h_a\right) \leq \frac{1}{1-\mu_a}\left(\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})\big] - \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})^2\big] + \delta_{cal}\right) \tag{9}$$

$$\frac{1}{\mu_a}\left(\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})\big] - \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})^2\big] - \delta_{cal}\right) \leq c_{fn}\left(h_a\right) \leq \frac{1}{\mu_a}\left(\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})\big] - \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})^2\big] + \delta_{cal}\right) \tag{10}$$

where $c_{err}\left(h_a\right)$, $c_{fp}\left(h_a\right)$, and $c_{fn}\left(h_a\right)$ represent the generalized error rate, generalized false-positive rate (FPR), and generalized false-negative rate (FNR). μ_a is the base rate for group G_a and $h_a(\mathcal{O}_\mathrm{T})$ is the classifier for group G_a.

Proof. First, for the generalized error rate we note from Definition 1 that

$$\begin{aligned} c_{err}\left(h_a\right) &= \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[\,|\,h_a(\mathcal{O}_\mathrm{T}) - y|\big] \\ &= \int_0^1 p\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T}) = p \mid y = 0\big]\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[y = 0\big] + (1-p)\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T}) = p \mid y = 1\big]\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[y = 1\big]\,dp \end{aligned}$$

Now we apply Bayes rule to find

$$\begin{aligned} c_{err}\left(h_a\right) &= \int_0^1 \Big(p\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[y = 0 \mid h_a(\mathcal{O}_\mathrm{T}) = p\big] + (1-p)\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[y = 1 \mid h_a(\mathcal{O}_\mathrm{T}) = p\big]\Big)\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T}) = p\big]\,dp \\ &= \int_0^1 \Big(p\left(1 - \mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[y = 1 \mid h_a(\mathcal{O}_\mathrm{T}) = p\big]\right) + (1-p)\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[y = 1 \mid h_a(\mathcal{O}_\mathrm{T}) = p\big]\Big)\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T}) = p\big]\,dp \end{aligned} \tag{11}$$

Working out the first part of Equation (11)

$$\begin{aligned}
&\int_0^1 p\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[y=1 \mid h_a(\mathcal{O}_\mathrm{T})=p\big]\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})=p\big]dp \\
&\quad= \int_0^1 p(p+\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[y=1 \mid h_a(\mathcal{O}_\mathrm{T})=p\right]-p)\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})=p\right]dp \\
&\quad\leq \int_0^1 \left(p^2+|\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[y=1 \mid h_a(\mathcal{O}_\mathrm{T})=p\right]-p|\right)\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})=p\right]dp \\
&\quad\leq \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}[h_a(\mathcal{O}_\mathrm{T})^2]+\delta_{cal}
\end{aligned} \tag{12}$$

Similarly, we can work out the lower bound

$$\begin{aligned}
&\int_0^1 p\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[y=1 \mid h_a(\mathcal{O}_\mathrm{T})=p\right]\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})=p\right]dp \\
&\quad\geq \int_0^1 \left(p^2-|\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[y=1 \mid h_a(\mathcal{O}_\mathrm{T})=p\right]-p|\right)\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})=p\right]dp \\
&\quad\geq \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}[h_a(\mathcal{O}_\mathrm{T})^2]-\delta_{cal}
\end{aligned} \tag{13}$$

Working out the second part of Equation (11)

$$\begin{aligned}
&\int_0^1 (1-p)(\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[y=1 \mid h_a(\mathcal{O}_\mathrm{T})=p\right])\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})=p\big]dp \\
&\quad= \int_0^1 (1-p)(p+\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[y=1 \mid h_a(\mathcal{O}_\mathrm{T})=p\right]-p)\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})=p\big]dp \\
&\quad\leq \int_0^1 (p(1-p)+|\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[y=1 \mid h_a(\mathcal{O}_\mathrm{T})=p\right]-p|)\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})=p\big]dp \\
&\quad\leq \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}[h_a(\mathcal{O}_\mathrm{T})]-\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}[h_a(\mathcal{O}_\mathrm{T})^2]+\delta_{cal}
\end{aligned} \tag{14}$$

and the lower bound

$$\begin{aligned}
&\int_0^1 (1-p)(\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[y=1 \mid h_a(\mathcal{O}_\mathrm{T})=p\right])\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})=p\big]dp \\
&\quad= \int_0^1 (1-p)(p+\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[y=1 \mid h_a(\mathcal{O}_\mathrm{T})=p\right]-p)\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})=p\big]dp \\
&\quad\geq \int_0^1 (p(1-p)-|\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[y=1 \mid h_a(\mathcal{O}_\mathrm{T})=p\right]-p|)\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})=p\big]dp \\
&\quad\geq \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}[h_a(\mathcal{O}_\mathrm{T})]-\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}[h_a(\mathcal{O}_\mathrm{T})^2]-\delta_{cal}
\end{aligned} \tag{15}$$

Combining Equations (12) to (15) with Equation (11), we find

$$2\left(\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})\big]-\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})^2\big]-\delta_{cal}\right)\leq c_{err}\left(h_a\right)\leq 2\left(\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})\big]-\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[h_a(\mathcal{O}_\mathrm{T})^2\big]+\delta_{cal}\right)$$

Second, for the generalized false-negative rate we note from Definition 1 that:

$$\begin{aligned}
c_{fn}\left(h_t\right) &= \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\big[1-h_a(\mathcal{O}_\mathrm{T}) \mid y=1\big] \\
&= \int_0^1 (1-p)\ \mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})=p \mid y=1\right]dp
\end{aligned}$$

Applying Bayes rule

$$\begin{aligned}
c_{fn}\left(h_t\right) &= \int_0^1 (1-p)\ \frac{\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[y=1 \mid h_a(\mathcal{O}_\mathrm{T})=p\right]}{\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[y=1\right]}\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})=p\right]dp. \\
&= \frac{1}{\mu_t}\int_0^1 (1-p)\left(\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[y=1 \mid h_a(\mathcal{O}_\mathrm{T})=p\right]\right)\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})=p\right]dp.
\end{aligned}$$

Substituting Equations (14) and (15) we find

$$\frac{1}{\mu_a}\left(\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})\right] - \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})^2\right] - \delta_{cal}\right) \leq c_{fp}\left(h_a\right) \leq \frac{1}{\mu_a}\left(\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})\right] - \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})^2\right] + \delta_{cal}\right)$$

Finally, for the generalized false-positive rate we note from Definition 1 that

$$\begin{aligned}
c_{fp}\left(h_a\right) &= \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T}) \mid y=0\right] \\
&= \int_0^1 p\ \mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T}) = p \mid y=0\right] dp \\
&= \int_0^1 p\ \frac{1-\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[y=1 \mid h_a(\mathcal{O}_\mathrm{T})=p\right]}{1-\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[y=1\right]}\,\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})=p\right] dp \\
&= \frac{1}{1-\mu_q}\int_0^1 p\left(1-\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[y=1 \mid h_a(\mathcal{O}_\mathrm{T})=p\right]\right)\mathrm{P}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})=p\right] dp
\end{aligned}$$

Substituting Equations (12) and (13) we find

$$\frac{1}{1-\mu_a}\left(\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})\right] - \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})^2\right] - \delta_{cal}\right) \leq c_{fn}\left(h_a\right) \leq \frac{1}{1-\mu_a}\left(\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})\right] - \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})^2\right] + \delta_{cal}\right)$$

The respective error rates for the perfectly calibrated case can be directly obtained by dropping δ_{cal} from Equations (8) to (9)

$$c_{err}\left(h_a\right) = 2\left(\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})\right] - \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})^2\right]\right) \tag{16}$$

$$c_{fp}\left(h_a\right) = \frac{1}{1-\mu_a}\left(\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})\right] - \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})^2\right]\right) \tag{17}$$

$$c_{fn}\left(h_a\right) = \frac{1}{\mu_a}\left(\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})\right] - \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})^2\right]\right) \tag{18}$$

□

Confidence thresholds

There are two confidence thresholds that, when reached, stop further feature acquisition. The classifier starts with an empty feature set at $h_a(\emptyset) = \mu_a$. The effective thresholds are therefore always $\alpha_u > \mu_a$ and the lower threshold $\alpha_l < \mu_a$. When the classifier stops at $t = T$, it will find either $h_a(\mathcal{O}_\mathrm{T}) = \alpha_u$ or $h_a(\mathcal{O}_\mathrm{T}) = \alpha_l$. Applying these thresholds we find

$$\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})\right] - \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})^2\right] = \frac{1}{|G_a|}\sum_{\mathbf{x},y\sim G_a}\left(p_u(\alpha_u - \alpha_u^2) + p_l(\alpha_l - \alpha_l^2)\right)$$

where p_u is the probability that an individual stops at α_u and $p_l = 1 - p_u$ the probability that an individual stops at α_l. Also, we set $\alpha_l = 1 - \alpha_u$. Therefore

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})\right] - \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\mathrm{G_a}}\left[h_a(\mathcal{O}_\mathrm{T})^2\right] &= \frac{1}{|G_a|}\sum_{\mathbf{x},y\sim G_a}\left(p_u(\alpha_u - \alpha_u^2) + (1-p_u)((1-\alpha_u) - (1-\alpha_u)^2)\right) \\
&= \frac{1}{|G_a|}\sum_{\mathbf{x},y\sim G_a}\left(\alpha_u - \alpha_u^2\right)
\end{aligned}$$

We now set a target generalized error rate $c_{err}\left(h_a(\mathcal{O}_\mathrm{T})\right) = \beta_{err}$, equal for each subgroup a, that determines when to stop selecting additional features for each individual. We first derive the thresholds for the perfectly calibrated classifier in Equation (16)

$$\beta_{err} = 2\left(\alpha_u - \alpha_u^2\right) \qquad \beta_{err} = 2\left((1-\alpha_l) - (1-\alpha_l)^2\right)$$

$$\alpha_{u,err} = \frac{1}{2} + \frac{1}{2}\sqrt{1-2\beta_{err}} \qquad \alpha_{l,err} = \frac{1}{2} - \frac{1}{2}\sqrt{1-2\beta_{err}}$$

Similarly, using Equations (17) and (18), we find, when setting the target generalized FPR rate β_{fp} and the target generalized FNR rate β_{fn}, the thresholds

$$\alpha_{u,fp} = \frac{1}{2} + \frac{1}{2}\sqrt{1-4\beta_{fp}(1-\mu_a)} \qquad \alpha_{l,fp} = \frac{1}{2} - \frac{1}{2}\sqrt{1-2\beta_{fp}(1-\mu_a)}$$

$$\alpha_{u,fn} = \frac{1}{2} + \frac{1}{2}\sqrt{1-4\beta_{fn}\mu_a} \qquad \alpha_{l,fn} = \frac{1}{2} - \frac{1}{2}\sqrt{1-2\beta_{fn}\mu_a}$$

Generalizing these thresholds to the approximately calibrated case using Equations (8) to (10), we find that these thresholds will result in

$$\begin{aligned}\beta_{err} - \delta_{cal} \leq c_{err}(h_a) \leq \beta_{err} + \delta_{cal} \\ \beta_{fp} - \delta_{cal} \leq c_{fp}(h_a) \leq \beta_{fp} + \delta_{cal} \\ \beta_{fn} - \delta_{cal} \leq c_{fn}(h_a) \leq \beta_{fn} + \delta_{cal}\end{aligned}$$

Confidence thresholds for arbitrary cost functions

In the main text, we formulate confidence thresholds for two distinct cases. First, we formulate them for the case where true positives and true negatives are equally desirable and one thus aims to equalize the accuracy across subgroups. Second, we formulate them for a case where an individual cares about equal false negatives rates. Here we show that the confidence thresholds generalize to an arbitrary cost function z_a, a linear function in $c_{fp}(h_a)$ and $c_{fn}(h_a)$

$$z_a(h_a(\mathcal{O}_t)) = b_a c_{fp}(h_a(\mathcal{O}_t)) + c_a c_{fn}(h_a(\mathcal{O}_t)) \tag{19}$$

with $b_a + c_a = 1$. We can reformulate this z_a using the generalized definitions in Definition 1

$$z_a(h_a(\mathcal{O}_t)) = b_a \frac{1}{|G_a|(1-\mu_t)} \sum_{(\mathbf{x},y)\in G_a} \mathbb{1}_{y=0} h_a(\mathcal{O}_t) + c_a \frac{1}{|G_a|\mu_t} \sum_{(\mathbf{x},y)\in G_a} \mathbb{1}_{y=1}(1 - h_a(\mathcal{O}_t)) \tag{20}$$

where we normalize by the total number of negative individuals $|G_a|(1-\mu_t)$ for c_{fp} and the number of positive individuals $|G_a|\mu_t$ for c_{fn}. Replacing the ground truth labels with the probabilistic estimates:

$$z_a(h_a(\mathcal{O}_t)) = \left(\frac{b_a}{|G_a|(1-\mu_t)} + \frac{c_a}{|G_a|\mu_t}\right) \sum_{(\mathbf{x},y)\in G_a} h_a(\mathcal{O}_t)(1 - h_a(\mathcal{O}_t)) \tag{21}$$

Now we define a target cost function β_z to find the stopping criteria for each group such that $\mathbb{E}_{(\mathbf{x},y)\sim G_a}[z_a(\mathcal{O}_T)] = \beta_z$ for all groups a and isolate $h_a(\mathcal{O}_T)$. We find a set of confidence thresholds that ensure

$$\beta_z = \left(\frac{b_a}{1-\mu_t} + \frac{c_a}{\mu_t}\right) h_a(\mathcal{O}_t)(1 - h_a(\mathcal{O}_t)) \tag{22}$$

which leads to the stopping criteria

$$\alpha_u = \frac{1}{2} + \frac{1}{2}\sqrt{1 - 4\beta_z \left(\frac{b_a}{|G_a|(1-\mu_t)} + \frac{c_a}{|G_a|\mu_t}\right)^{-1}} \tag{23}$$

$$\alpha_l = \frac{1}{2} - \frac{1}{2}\sqrt{1 - 4\beta_z \left(\frac{b_a}{|G_a|(1-\mu_t)} + \frac{c_a}{|G_a|\mu_t}\right)^{-1}} \tag{24}$$

Naturally, when $b_a = 0$ and $c_a = 1$ we retrieve the thresholds we found for achieving equal false negative rates, while when $b_a = 1$ and $c_a = 0$ we retrieve the thresholds for achieving equal false positive rates.

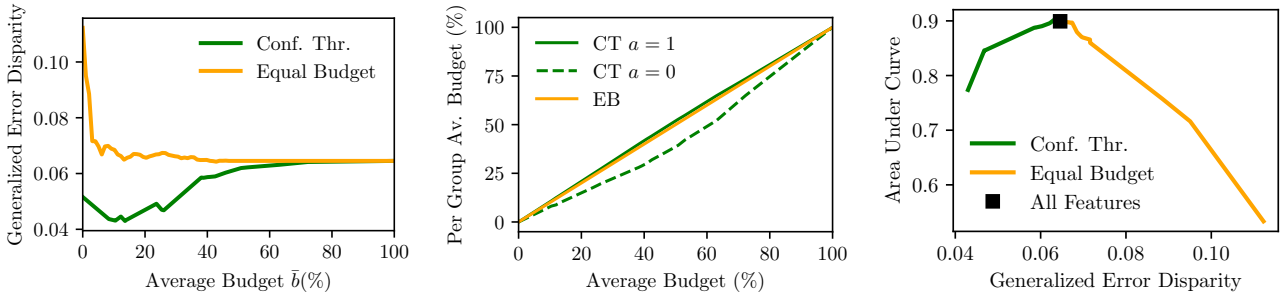

Figure SM1: Confidence thresholds for achieving equal generalized error rates across the White and Non-white subgroups in the Adult Income dataset. In each plot, the curves are generated by sweeping the target error rate β_{err}, effectively changing the average budget used across groups. Left: the residual generalized error disparity for the confidence thresholds (blue) and the equal budget benchmark (red). Center: The average budget per group versus the total average budget. The dashed and solid blue lines represent the average budget used for respectively the White ($a = 1$) and Non-white ($a = 0$) subgroups while the red line represents the average budget for both groups when using the benchmark method. Right: The Pareto front of the AUC versus disparity trade-off for our method and the benchmark method as well as for the classifier with access to all features in black.

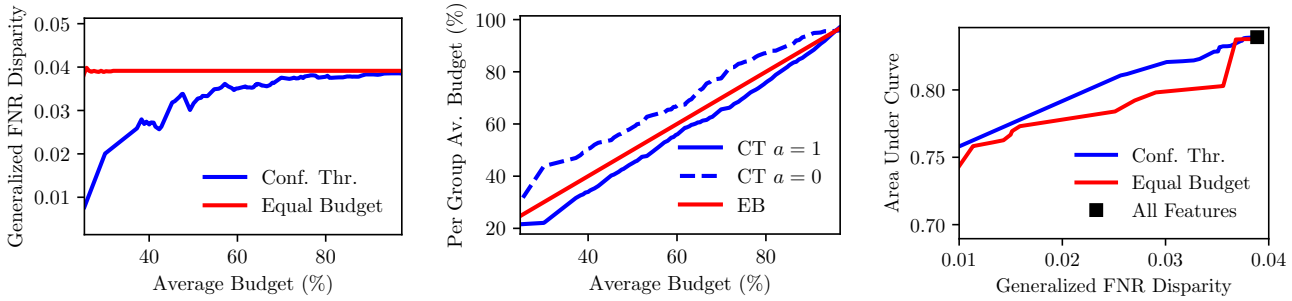

Figure SM2: Confidence thresholds for achieving equal generalized FNR rates across the Urban and Rural subgroups in the Mexican Poverty dataset. In each plot, the curves are generated by sweeping the target error rate β_{fnr}. Left: the residual generalized FNR disparity for the confidence thresholds (blue) and the equal budget benchmark (red). Center: The average budget per group versus the total average budget across groups. The dashed and solid blue lines represent the average budget used for respectively the Urban ($a = 1$) and Rural ($a = 0$) subgroups while the red line represents the average budget for both groups when using the benchmark method. Right: The Pareto front of the AUC-disparity trade-off for our method and the benchmark method. The black square represents the classifier that has access to all features.

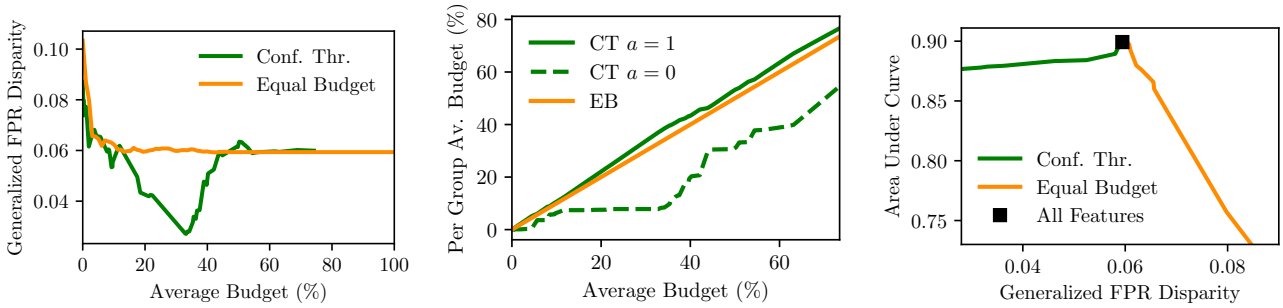

Figure SM3: Confidence thresholds for achieving equal generalized FPR rates across the White and Non-white subgroups in the Adult Income dataset. In each figure, the curves are generated by sweeping the target error rate β_{fpr}. Left: the residual generalized FPR disparity for the confidence thresholds (blue) and the equal budget benchmark (red). Center: The average budget per group versus the total average budget across groups. The dashed and solid blue lines represent the average budget used for respectively the White ($a = 1$) and Non-white ($a = 0$) subgroups while the red line represents the average budget for both groups when using the benchmark method. Right: The Pareto front of the AUC-disparity trade-off for our method and the benchmark method. The black square represents the classifier that has access to all features.