Fair Selective Classification via Sufficiency

Anonymous Authors[1]

Abstract

Selective classification is a powerful tool for decision-making in scenarios where mistakes are costly but abstentions are allowed. In general, by allowing a classifier to abstain, one can improve the performance of a model at the cost of reducing coverage and classifying fewer samples. However, recent work has shown, in some cases, that selective classification can magnify disparities between groups, and has illustrated this phenomenon on multiple real-world datasets. We prove that the sufficiency criterion can be used to mitigate these disparities by ensuring that selective classification increases performance on all groups, and introduce a method for mitigating the disparity in precision across the entire coverage scale based on this criterion. We then provide an upper bound on the conditional mutual information between the class label and sensitive attribute, conditioned on the learned features, which can be used as a regularizer to achieve fairer selective classification. The effectiveness of the method is demonstrated on the Adult, CelebA, Civil Comments, and CheXpert datasets.

1. Introduction

As machine learning applications continue to grow in scope and diversity, its use in many industries raises increasingly many ethical and legal concerns, especially those of fairness and bias in predictions made by automated systems (Selbst et al., 2019; Bellamy et al., 2018; Meade, 2019). As systems are trusted to aid or make decisions regarding loan applications, criminal sentencing, and even health care, it is more important than ever that these predictions be free of bias.

The field of fair machine learning is rich with both problems and proposed solutions, aiming to provide unbiased decision systems for various applications. A number of different definitions and criteria for fairness have been proposed, as well as a variety of settings where fairness might be applied.

One major topic of interest in fair machine learning is that of fair classification, whereby we seek to make a classifier "fair" for some definition of fairness that varies according to the application. In general, fair classification problems arise when we have *protected groups* that are defined by a shared *sensitive attribute* (e.g. race, gender), and we wish to ensure that we are not biased against any one group with the same sensitive attribute.

In particular, one sub-setting of fair classification which exhibits an interesting fairness-related phenomenon is that of selective classification. Generally speaking, selective classification is a variant of the classification problem where a model is allowed to *abstain* from making a decision. This has applications in settings where making a mistake can be very costly, but abstentions are not (e.g. if the abstention results in deferring classification to a human actor).

In general, selective classification systems work by assigning some measure of confidence about their predictions, and then deciding whether or not to abstain based on this confidence, usually via thresholding.

The desired outcome is obvious: the higher the confidence threshold for making a decision (i.e. the more confident one needs to be to not abstain), the lower the *coverage* (proportion of samples for which a decision is made) will be. But in return, one should see better performance on the remaining samples, as one is only making decisions when one is very sure of the outcome. In practice, for most datasets, with the correct choice of confidence measure and the correct training algorithm, this outcome is observed.

However, recent work has revealed that selective classification can magnify disparities between groups as the coverage decreases, even as overall performance increases. (Jones et al., 2020). This, of course, has some very serious consequences for systems that require fairness, especially if it appears at first that predictions are fair enough under full coverage (i.e. when all samples are being classified).

Thus, we seek a method for enforcing fairness which ensures that a classifier is fair even if it abstains from classifying on a large number of samples. In particular, having a measure of confidence that is reflective of accuracy for each group

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

can ensure that thresholding doesn't harm one group more than another. This property can be achieved by applying a condition known as *sufficiency*, which ensures that our predictive scores in each group are such that they provide the same accuracy at each confidence level (Barocas et al., 2019). This condition also ensures that the precision on all groups increases when we apply selective classification, and can help mitigate the disparity between groups as we decrease coverage.

The sufficiency criteria can be formulated as enforcing a conditional independence between the label and sensitive attribute, conditioned on the learned features, and thus allows for a relaxation and optimization method that centers around the mutual information. However, to impose this criteria, we require the use of a penalty term that includes the conditional mutual information between two discrete or continuous variables conditioned on a third continuous variable. Existing method for computing the mutual information for the purposes of backpropagation tend to struggle when the term in the condition involves the learned features. In order to facilitate this optimization, we thus derive an upper-bound approximation of this quantity.

In this paper, we make two main contributions. Firstly, we prove that sufficiency can be used to train fairer selective classifiers which ensure that precision always increases as coverage is decreased for all groups. Secondly, we derive a novel upper bound of the conditional mutual information which can be used as a regularizer to enforce the sufficiency criteria, then show that it works to mitigate the disparities on real-world datasets.

2. Background

2.1. The Fair Classification Problem

We begin with the standard supervised learning setup of predicting the value of a target variable $Y \in \mathcal{Y}$ using a set of decision or predictive variables $X \in \mathcal{X}$ with training samples $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. For example, X may be information about an individual's credit history, and Y is whether the individual will pay back a certain loan. In general, we wish to find features $\Phi(x) \in \mathbb{R}^{d_\Phi}$, which are predictive about Y, so that we can construct a good predictor $\hat{y} = T(\Phi(x))$ of y under some loss criteria $L(\hat{y}, y)$.

Now suppose we have some sensitive attributes $D \in \mathcal{D}$ we wish to be "fair" about (e.g. race, gender), and training samples $\{(x_1, y_1, d_1), \ldots, (x_n, y_n, d_n)\}$. For example, in the banking system, predictions about the chance of someone repaying a loan (Y) given factors about one's financial situation (X) should not be determined by gender (D). While D can be continuous or discrete (and our method generalizes to both cases), we focus in this paper on the case where D is discrete, and refer to members which share the same value of D as being in the same *group*. This allows us to formulate metrics based on group-specific performance.

There are numerous metrics and criteria for what constitutes a fair classifier, many of which are mutually exclusive with one another outside of trivial cases. One important criteria is *positive predictive parity* (Corbett-Davies et al., 2017; Pessach & Shmueli, 2020), which is satisfied when the precision (which we denote as PPV, after Positive Predictive Value) for each group is the same, that is:

$$\forall a, b \in \mathcal{D}, \mathbb{P}(Y=1|\hat{Y}=1, D=a) = \mathbb{P}(Y=1|\hat{Y}=1, D=b). \tag{1}$$

This criteria is especially important in applications where false positives are particularly harmful (e.g. criminal sentencing or loan application decisions) and having one group falsely labeled as being in the positive group could lead to great harm or contribute to further biases. Looking at precision rates can also reveal disparities that may be hidden by only considering the differences in accuracies across groups (Angwin et al., 2016).

When D is binary, an intuitive way to measure the severity of violations of this condition is to measure the difference in precision between the two groups:

$$\Delta_{PPV} \triangleq \mathbb{P}(Y=1|\hat{Y}=1, D=0) - \mathbb{P}(Y=1|\hat{Y}=1, D=1). \tag{2}$$

A number of methods exist for learning fairer classifiers. (Calders et al., 2009) and (Menon & Williamson, 2018) reweight probabilities to ensure fairness, while (Zafar et al., 2017) proposes using covariance-based constraints to enforce fairness criteria, and (Zhang et al., 2018) uses an adversarial method, requiring the training of an adversarial classifier. Still others work by directly penalizing some specific fairness measure (Zemel et al., 2013a; Calmon et al., 2017).

Many of these methods also work by penalizing some approximation or proxy for the mutual information. (Mary et al., 2019; Baharlouei et al., 2019; Lee et al., 2020) propose the use of the HGR maximal correlation as a regularizer for the independence and separation constraints (which requires that $\hat{Y} \perp D$ and $\hat{Y} \perp D \mid Y$, respectively), which has been shown to be an approximation for the mutual information (Huang et al., 2019). Finally, (Cho et al., 2020) approximates the mutual information using a variational approximation. However, none of these methods are designed to tackle the selective classification problem.

2.2. Selective Classification

In selective classification, a predictive system is given the choice of either making a prediction $\hat{Y}$ or abstaining from the decision. The core assumption underlying selective classification is that there are samples for which a system is

more confident about its prediction, and by only making predictions when it is confident, the performance will be improved. To enable this, we must have a *confidence score* $\kappa(x)$ which represents the model's certainty about its prediction on a given sample x (Geifman & El-Yaniv, 2017). Then, we threshold on this value to decide whether to make a decision or to abstain. We define the *coverage* as the fraction of samples for which we do not abstain on (i.e. the fraction of samples that we make predictions on).

As is to be expected, when the confidence is a good measure of the probability of making a correct prediction, then as we increase the minimum confidence threshold for making the prediction (thus decreasing the coverage), we should see the risk on the classified samples decrease or the accuracy over the classified samples increase. This leads us to the *accuracy-coverage* tradeoff, which is central to selective classificaton (though we note here the warning from the previous section about accuracy not telling the whole story).

Selective classifiers can work *a posteriori* by taking in an existing classifier and deriving an uncertainty measure from it for which to threshold on (Geifman & El-Yaniv, 2017), or a selective classifier can be trained with an objective that is designed to enable selective classificaiton (Cortes et al., 2016; Yildirim et al., 2019).

One common method of extracting a confidence score from an existing network is to take the softmax response $s(x)$ as a measure of confidence. In the case of binary classifcation, to better visualize the distribution of the scores, we define the confidence using a monotonic mapping of $s(x)$:

$$\kappa = \frac{1}{2}\log\left(\frac{s(x)}{1-s(x)}\right) \tag{3}$$

which maps $[0.5, 1]$ to $[0, \infty]$ and provides much higher resolution on the values close to 1.

Finally, to measure the effectiveness of selective classification, we can plot the accuracy-coverage curve, and then compute the area under this curve to encapsulate the performance across different coverages (Franc & Prusa, 2019).

2.3. Biases in Selective Classification

(Jones et al., 2020) has shown that in some cases, when coverage is decreased, the difference in recall between groups can sometimes increase, magnifying disparities between groups and increasing unfairness. In particular, they have shown that in the case of the CelebA and CivilComments dataset, decreasing the coverage can also decrease the recall on the worst-case group.

In general, this phenomenon occurs due to a difference between the average margin distribution and the group-specific margin distributions, resulting in different levels of performance when thresholding, as illustrated in Figure 1.

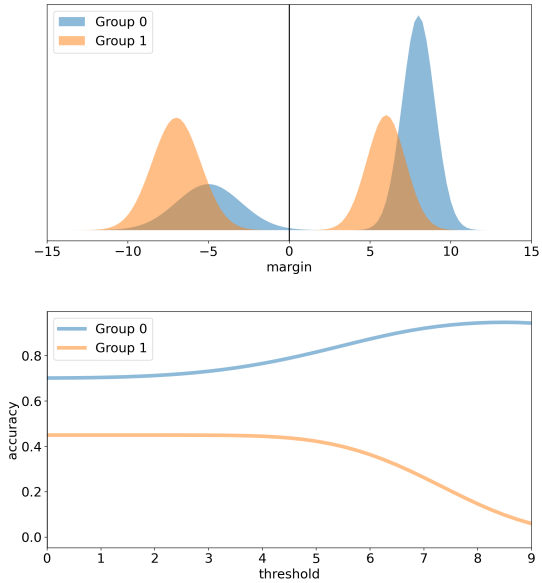


Figure 1. (Top) When margin distributions are not aligned, (Bottom) then as we sweep over the threshold τ, the accuracies for the groups do not necessarily move in concert with one another.

The *margin* M of a classifier is defined as $\kappa(x)$ when $\hat{y}(x) = y$ and $-\kappa(x)$ otherwise. If we let τ be our threshold, then a selective classifier makes the correct prediction when $M(x) \geq \tau$ and incorrect predictions when $M(x) \leq -\tau$. We also denote its probability density function (PDF) and cumulative density function (CDF) as f_M and F_M, respectively. Then, the selective accuracy is

$$A_F(\tau) = \frac{1 - F_M(\tau)}{F_M(-\tau) + 1 - F_M(\tau)} \tag{4}$$

for a given threshold. We can analogously compute the selective precision by conditioning on $\hat{Y} = 1$,

$$PPV_F(\tau) = \frac{1 - F_{M|\hat{Y}=1}(\tau)}{F_{M|\hat{Y}=1}(-\tau) + 1 - F_{M|\hat{Y}=1}(\tau)}. \tag{5}$$

We can also analogously define the distributions of the margin for each group using $f_{M,D}$ and $F_{M,d}$ for group $d \in \mathcal{D}$.

(Jones et al., 2020) proposes a number of different situations for which average accuracy could increase but worst-group accuracy could decrease based on their relative margin distributions. For example, if F is left-log-concave (e.g. Gaussian), then $A_F(\tau)$ is monotonically increasing when $A_F(0) \geq 0.5$ and monotonically decreasing otherwise. Thus, if $A_F(0) > 0.5$ but $A_{F_d}(0) < 0.5$, then average accuracy may increase as we increase τ (and thus decrease coverage) but the accuracy on group d may decrease, thus resulting in magnified disparity. This same phenomenon occurs with the precision when we condition on $\hat{Y} = 1$. In

general, when margin distributions are not aligned between groups, disparity can increase as one sweeps over the threshold τ. Further subdividing groups according to their label yields the difference in recall rates observed.

3. Fair Selective Classification with Sufficiency

3.1. Sufficiency and Fair Selective Classification

Our solution to the fair selective classification problem is to apply the sufficiency criteria to the learned features.

Sufficieny requires that $Y \perp D \mid \hat{Y}$ or $Y \perp D \mid \Phi(X)$, i.e., the prediction completely subsumes all information about the sensitive attribute that is relevant to the label (Cleary, 1966). When Y is binary, the sufficiency criteria requires that (Barocas et al., 2019):

$$\mathbb{P}(Y = 1|\Phi(x), D = a) = \mathbb{P}(Y = 1|\Phi(x), D = b) \quad \forall a, b \in \mathcal{D}. \tag{6}$$

The application of this criteria to fair selective classification comes to us by way of *Calibration by Group*. Calibration by group requires that there exists a score function $R = s(x)$ such that, for all $r \in (0, 1)$: (Chouldechova, 2017):

$$\mathbb{P}(Y = 1|R = r, D = a) = r \ \ \forall a \in \mathcal{D}. \tag{7}$$

The following result from (Barocas et al., 2019) links calibration and sufficiency:

Theorem 1. *If a classifier has sufficient features Φ, then there exists a mapping $h(\Phi) : \mathbb{R}^{d_\Phi} \to [0, 1]$ such that $h(\Phi)$ is calibrated by group.*

If we can find sufficient features $\Phi(X)$, so that the score function is calibrated by group based these features, then we have the following result (the proof can be found in the Appendices):

Theorem 2. *If a classifier has a score function $R = s(x)$ which is calibrated by group, and selective classification is performed using confidence κ as defined in (3), then for all groups $d \in \mathcal{D}$ we have that both $A_F(\tau)$ and $PPV_{F_d}(\tau)$ are **monotonically increasing** with respect to τ. Furthermore, we also have that $A_{F_d}(0) > 0.5$ and $PPV_{F_d}(0) > 0.5$.*

From this, we can guarantee that as we sweep through the threshold, we will never penalize performance of any one group in service of increasing the overall precision. Furthermore, in most real-world applications, the precision on the best-performing groups tends to saturate very quickly to values close to 1 when coverage is reduced, and thus, if we can guarantee that the precision increases on the worst performing group as well, then in general, the difference in precision between groups decreases as coverage decreases.

3.2. Imposing the Sufficiency Condition

From the above theorem, we can see that a sufficient classifier should yield the desired property of enabling fair selective classification. It should ensure that as we sweep over the coverage, the performance of one group is not penalized in the service of improving the performance of another group or improving the average performance.

In order to impose sufficiency as a fairness criteria, we formulate the following training objective as a constrained optimization of a standard loss function:

$$\begin{aligned} \min_{\theta} \quad & L(\hat{y}, y) \\ \text{s.t.} \quad & Y \perp D|\Phi(X), \end{aligned} \tag{8}$$

where $\hat{y} = T(\Phi(x))$, and θ are the model parameters for both Φ and T. One possible way of representing the sufficiency constraint is by using the mutual information:

$$\begin{aligned} \min_{\theta} \quad & L(\hat{y}, y) \\ \text{s.t.} \quad & I(Y; D|\Phi(X)) = 0. \end{aligned} \tag{9}$$

This follows from the fact that $Y \perp D \mid \Phi(X)$ is satisfied if and only if $I(Y; D|\Phi(X)) = 0$. This provides us with a simple relaxation of the constraint into the following form:

$$\min_{\theta} L(\hat{y}, y) + \lambda I(Y; D|\Phi(X)). \tag{10}$$

We note here that existing works using mutual information for fairness are ill-equipped to handle this condition, as they assume that it is not the features that will be conditioned on, but rather that the penalty will be the mutual information between the sensitive attribute and the features (e.g. penalizing $I(\Phi(X); D)$ for demographic parity), possibly conditioned on the label (e.g. penalizing $I(\Phi(X); D|Y)$ in the case of equalized odds). As such, existing methods either assume that the variable being conditioned on is discrete (Calmon et al., 2017; Zemel et al., 2013b; Hardt et al., 2016), become unstable when the features are placed in the condition (Mary et al., 2019), or simply do not allow for conditioning of this type due to their formulation (Grari et al., 2019; Baharlouei et al., 2019).

Thus, in order to approximate the mutual information for our purposes, we must first derive an upper bound for the mutual information which is computable in our applications. Our bound is inspired by the work of (Cheng et al., 2020) and is stated in the following theorem:

Theorem 3. *For random variables X, Y and Z, we have*

$$I_{\text{UB}}(X; Y|Z) \geq I(X; Y|Z), \tag{11}$$

where equality is achieved if and only if $X \perp Y \mid Z$, and

$$\begin{aligned} I_{\text{UB}}(X; Y|Z) \triangleq \ & \mathbb{E}_{P_{XYZ}}\left[\log P(Y|X, Z)\right] \\ & - \mathbb{E}_{P_X}\left[\mathbb{E}_{P_{YZ}}\left[\log P(Y|X, Z)\right]\right]. \end{aligned} \tag{12}$$

Proof. The conditional mutual information can be written as

$$I(X;Y|Z) = \mathbb{E}_{P_{XYZ}}\left[\log P(Y|X,Z)\right] - \mathbb{E}_{P_{YZ}}\left[\log P(Y|Z)\right]. \quad (13)$$

Thus,

$$I_{\mathrm{UB}}(X;Y|Z) - I(X;Y|Z) = \mathbb{E}_{P_{YZ}}\left[\log P(Y|Z) + \mathbb{E}_{P_X}\left[-\log P(Y|X,Z)\right]\right]. \quad (14)$$

Note that $-\log(\cdot)$ is convex,

$$\mathbb{E}_{P_X}\left[-\log P(Y|X,Z)\right] \geq -\log \mathbb{E}_{P_X}\left[P(Y|X,Z)\right] = -\log P(Y|Z), \quad (15)$$

which completes the proof. □

Thus, $I(Y;D|\Phi(X))$ can be upper bounded by I_{UB} as:

$$I(Y;D|\Phi(X)) \leq \mathbb{E}_{P_{XYD}}\left[\log P(Y|\Phi(X),D)\right] - \mathbb{E}_{P_D}\left[\mathbb{E}_{P_{XY}}\left[\log P(Y|\Phi(X),D)\right]\right]. \quad (16)$$

Since $P(y|\Phi(x),d)$ is unknown in practice, we need to use a variational distribution $q(y|\Phi(x),d;\theta)$ with parameter θ to approximate it. Here, we adopt a neural net that predicts Y based on feature $\Phi(X)$ and sensitive attribute D as our variational model $q(y|\Phi(x),d;\theta)$.

However, in many cases, X will be continuous, high-dimensional data (e.g. images), while D will be a discrete, categorical variable (e.g. gender, ethnicity), therefore, it would be more convenient to instead formulate the model as $q(y|\Phi(x);\theta_d)$, i.e., to train a *group-specific* model for each $d \in \mathcal{D}$ to approximate $P(y|\Phi(x),d)$, instead of treating D as a single input to the neural net.

Then, we can compute the first term of the upper bound as the negative cross-entropy of the training samples using the "correct" classifier for each group (group-specific loss), and the second term as the cross-entropy of the samples using a randomly-selected classifier (group-agnostic loss) drawn according to the marginal distribution P_D. Thus, by replacing all expectations in (16) with empirical averages, the regularizer is given by

$$L_R \triangleq \frac{1}{n}\sum_{i=1}^{n}\Big(\log q(y_i|\Phi(x_i);\theta_{d_i}) - \log q(y_i|\Phi(x_i);\theta_{\widetilde{d}_i})\Big), \quad (17)$$

where $\widetilde{d}_i$ are drawn i.i.d. from the marginal distribution P_D, and for $d \in \mathcal{D}$,

$$\theta_d = \arg\max_{\theta} \sum_{i:\ d_i=d} \log q(y_i|\Phi(x_i);\theta). \quad (18)$$

Let T denote a *joint classifier* over all groups which is used to make final predictions, such that $\hat{y} = T(\Phi(x))$, then the overall loss function is

$$\min_{\theta_T,\theta_\Phi} \frac{1}{n}\sum_{i=1}^{n}\Big(L\big(T(\Phi(x_i)),y_i\big) + \lambda \log q(y_i|\Phi(x_i);\theta_{d_i}) - \lambda \log q(y_i|\Phi(x_i);\theta_{\widetilde{d}_i})\Big). \quad (19)$$

In practice, we train our model by alternating between the fitting steps in (18) and feature updating steps in (19), and the overall training process is described in Algorithm 1 and Figure 2. Intuitively, by trying to minimize the difference between the log-probability of the output of the correct model and that of the randomly-chosen one, we are trying to enforce $\Phi(x)$ to have the property that all group-specific models trained on it will be the same; that is:

$$q(y|\Phi(x);\theta_a) = q(y|\Phi(x);\theta_b), \quad \forall a,b \in \mathcal{D}. \quad (20)$$

This happens when $P(Y|\Phi(X),D) = P(Y|\Phi(X))$, which implies the sufficiency condition $Y \perp D|\Phi(X)$.

Algorithm 1 Training with sufficiency-based regularizer

Data: Training samples $\{(x_1,y_1,d_1),\ldots,(x_n,y_n,d_n)\}$, $\{\widetilde{d}_1,\ldots,\widetilde{d}_n\}$, which are drawn i.i.d. from the empirical distribution $\hat{P}_D$
Initialize Φ, T (parameterized by θ_ϕ and θ_T, respectively) and θ_d with pre-trained model, and let n_d be the number of samples in group d.
Compute the following losses:
Group-specific losses $L_d = -\sum_{i:\ d_i=d} \log q(y_i|\Phi(x_i);\theta)$
Joint loss $L_0 = \frac{1}{n}\sum_{i=1}^{n} L\big(T(\Phi(x_i)),y_i\big)$
Regularizer loss L_R defined in (17) including both Group-specific loss and Group-agnostic loss
for *each training iteration* **do**
  for $d = 1,\ldots,|\mathcal{D}|$ **do** `// Fit group-specific models`
    for $j = 1,\ldots,M$ **do** `// For each batch`
      $\theta_d \leftarrow \theta_d - \frac{1}{n_d}\eta_d \nabla_\theta L_d$
    end
  end
  for $j = 1,\ldots,N$ **do** `// For each batch`
    $\theta_\phi \leftarrow \theta_\phi - \frac{1}{n}\eta_f \nabla_{\theta_\phi}(L_0 + \lambda L_R)$ `// Update feature extractor`
    $\theta_T \leftarrow \theta_T - \frac{1}{n}\eta \nabla_{\theta_T} L_0$ `// Update joint classifier`
  end
end

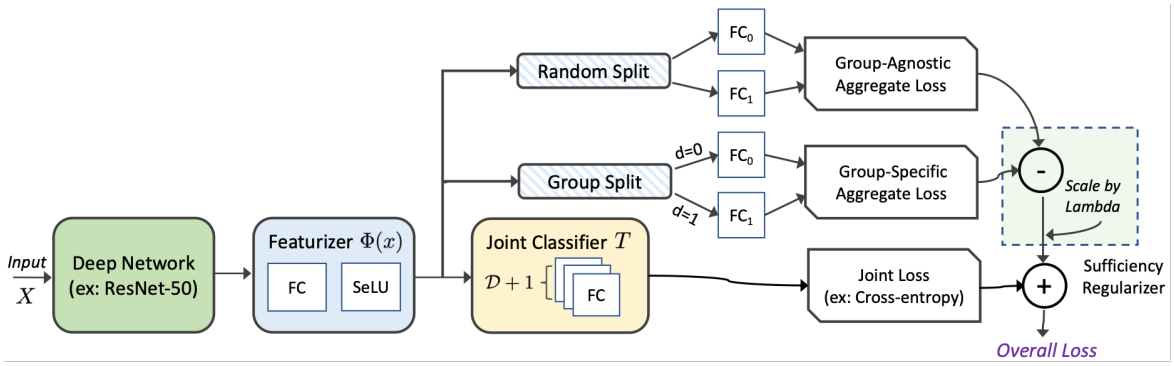


Figure 2. Diagram illustrating the computation of our sufficiency-based loss when D is binary.

Table 1. Summary of datasets.

Dataset	Modality	Target	Attribute
Adult	Demographics	Income	Sex
CelebA	Photo	Hair Colour	Gender
Civil Comments	Text	Toxicity	Christianity
CheXpert-device	X-ray	Disease	Support Device

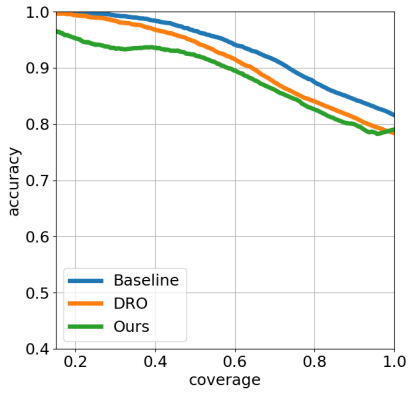


Figure 3. Overall accuracy-coverage curves for Adult dataset for the three methods.

4. Experimental Results

4.1. Datasets and Setup

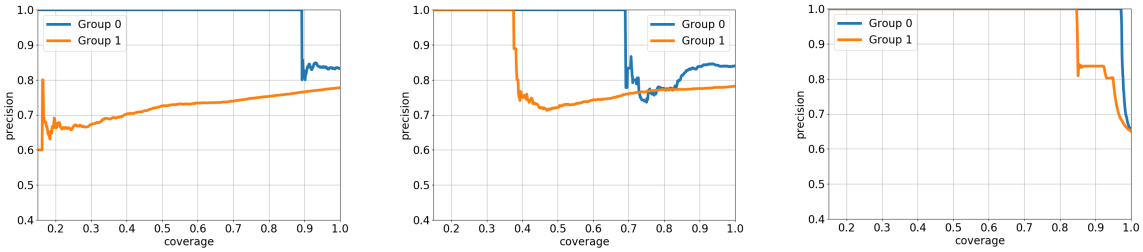
We test our method on four datasets which are commonly used in fairness: Adult[1], CelebA[2], Civil Comments[3], and CheXpert[4]. In all cases, we use the standard train/val/test splits packaged with the datasets and implemented our code in PyTorch. We set $\lambda = 0.7$ for all datasets as well, which we chose by sweeping over values of λ across all datasets.

The Adult dataset (Kohavi, 1996) consists of census data drawn from the 1994 Census database, with 48,842 samples. The data X consists of demographic information about individuals, including age, education, marital status, and country of origin. Following (Bellamy et al., 2018), we one-hot encode categorical variables and designate the binary-quantized income to be the target label Y and sex to be the sensitive attribute D. In order to simulate the bias phenomenon discussed in Section 2.3, we also drop all but the first 50 samples for which $D = 0$ and $Y = 1$. We then use a two-layers neural network with 80 nodes in the hidden layer for classification, as in (Mary et al., 2019), with the first layer serving as the feature extractor and the second as the classifier, and trained the network for 20 epochs.

The CelebA dataset (Liu et al., 2015) consists of 202,599 images of 10,177 celebrities, along with a list of attributes associated with them. As in (Jones et al., 2020), we use the images as our data X (resized to 224x224), the hair color (blond or not) as the target label Y, and the gender as the sensitive attribute D, then train a ResNet-50 model (He et al., 2016) (with initialization using pre-trained ImageNet weights) for 10 epochs on the dataset, with the penultimate layer as the feature extractor and the final layer as the classifier.

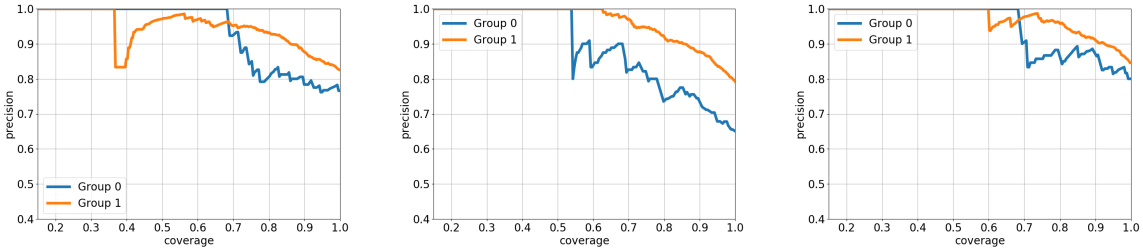
The Civil Comments dataset (Borkan et al., 2019) is a text-based dataset consisting of a collection of online comments, numbering 1,999,514 in total, on various news articles, along with metadata about the commenter and a label indicating whether the comment displays toxicity or not. As in (Jones et al., 2020), we let X be the text of the com-

[1]https://archive.ics.uci.edu/ml/datasets/adult
[2]http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
[3]https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data
[4]https://stanfordmlgroup.github.io/competitions/chexpert



(a) Baseline (b) DRO (c) Sufficiency-regularized

Figure 4. Group-specific precision-coverage curves for Adult dataset for the three methods.



(a) Baseline (b) DRO (c) Sufficiency-regularized

Figure 5. Group-specific precision-coverage curves for CheXpert dataset for the three methods.

ment, Y be the toxicity binary label, and D to be mention of Christianity. We pass the data first through a BERT model (Devlin et al., 2018) with Google's pre-trained parameters (Turc et al., 2019) and treat the output features as the input into our system. We then apply a 2-layer neural network to the BERT output with 80 nodes in the hidden layer, once again treating the layers as feature extractor and classifier, respectively. We trained the model for 20 epochs.

The CheXpert dataset (Irvin et al., 2019) comprises of 224,316 chest radiograph images from 65,240 patients with annotations for 14 different lung diseases. As in (Jones et al., 2020), we consider the binary classification task of detecting Pleural Effusion (PE). We set X to be the X-ray image of resolution 224x224, Y is whether the patient has PE, and D is the presence of a support device. We train a model by fine-tuning the DenseNet-121 (Huang et al., 2017) (with initialization using pre-trained ImageNet weights) for 10 epochs on the dataset, with the penultimate layer as the feature extractor and the final layer as the classifier.

We compared our results to a baseline where we only optimize the cross-entropy loss, as in standard classification. We also compared our method to the group DRO method of (Sagawa et al., 2019), using the code provided publicly on Github[5], which has been shown to mitigate the disparity in recall rates between groups in selective classification (Jones et al., 2020).

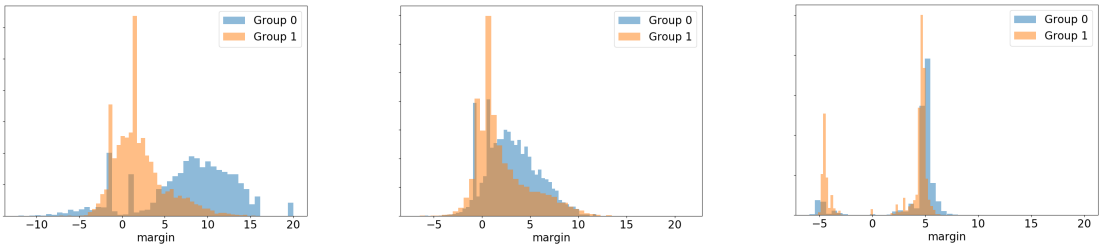
[5]https://github.com/kohpangwei/group_DRO

4.2. Results and discussion

Figure 3 shows the overall accuracy vs. coverage graphs for each method on the Adult dataset. We can see that, in all cases, selective classification increases the overall accuracy on the dataset, as is to be expected.

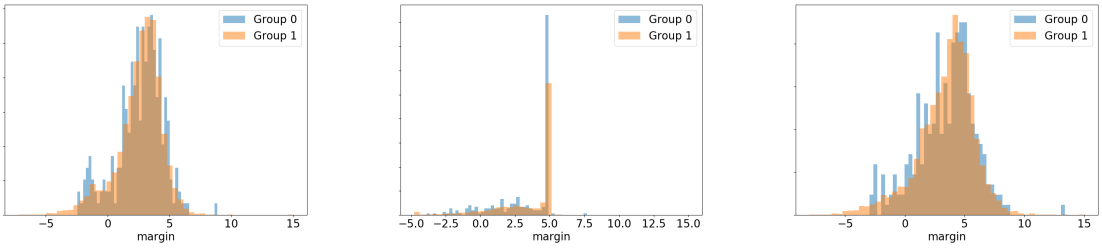
However, when we look at the group-specific precisions in Figure 4, we observe that, for the baseline method, this increase in performance comes at the cost of worse performance on the worst-case group. This phenomenon is heavily mitigated in the case of DRO, but there is still a gap in performance in the mid-coverage regime. Finally, our method shows the precisions converging to equality as coverage decreases very quickly. This can be explained by looking at the margin distributions for each method. The margin distribution histograms are plotted in Figure 6. We can see that the margin distributions are mismatched for the two groups in the baseline and DRO cases, but aligned for our sufficiency-based method.

Figure 5 and 7 show the group precisions and margin distributions for the CheXpert dataset. We can see that our method produces a smaller gap in precision at almost all coverages compared to the other two methods, and improves the worst-group precision. Note, in this use-case the presence of a support device (e.g., chest tubes) is spuriously correlated to being diagnosed as having PE (Oakden-Rayner et al., 2020). Thus, the worst-case group includes X-rays with a support device, that are diagnosed as PE negative.



(a) Baseline (b) DRO (c) Sufficiency-regularized

Figure 6. Margin distributions for Adult dataset for the three methods.



(a) Baseline (b) DRO (c) Sufficiency-regularized

Figure 7. Margin distributions for CheXpert dataset for the three methods.

Finally, in order to numerically evaluate the relative performances of the algorithms for all the datasets, we compute the following quantities: area under the average accuracy-coverage curve (Franc & Prusa, 2019) and area under the absolute difference in precision-coverage curve (or area between the precision-coverage curve for the two groups). Table 2 shows the results for each method and dataset.

From this, it is clear that while our method may incur a small decrease in overall accuracy in some cases, it reduces the disparity between the two groups, as desired.

5. Conclusion

Fairness in machine learning has never been a more important goal to pursue, and as we continue to root out the biases that plague our systems, we must be ever-vigilant of settings and applications where fairness techniques may need to be applied. We have introduced a method for enforcing fairness in selective classification, using a novel application of a novel bound for the conditional mutual information. And yet, the connection to mutual information suggests that there may be some grander picture yet to be seen, whereby the various mutual information-inspired methods may be unified. A central perspective on fairness grounded in such a fundamental quantity could prove incredibly insightful, both for theory and practice.

Table 2. Area under curve results for all datasets.

Dataset	**Method**	**Area under accuracy curve**	**Area between precision curves**
Adult	Baseline	0.931	0.220
	DRO	0.911	0.116
	Ours	0.887	0.021
CelebA	Baseline	0.852	0.094
	DRO	0.965	0.018
	Ours	0.975	0.013
Civil Comments	Baseline	0.888	0.026
	DRO	0.944	0.013
	Ours	0.943	0.010
CheXpert-device	Baseline	0.929	0.064
	DRO	0.933	0.080
	Ours	0.934	0.031

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. How we analyzed the compas recidivism algorithm. *ProPublica*, 2016. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Baharlouei, S., Nouiehed, M., Beirami, A., and Razaviyayn, M. Rènyi fair inference. *arXiv preprint arXiv:1906.12005*, 2019.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., et al. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, 2019.

Calders, T., Kamiran, F., and Pechenizkiy, M. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 13–18. IEEE, 2009.

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pp. 3992–4001, 2017.

Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z., and Carin, L. Club: A contrastive log-ratio upper bound of mutual information. In *International Conference on Machine Learning*, pp. 1779–1788. PMLR, 2020.

Cho, J., Hwang, G., and Suh, C. A fair classifier using mutual information. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2521–2526. IEEE, 2020.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Cleary, T. A. Test bias: Validity of the scholastic aptitude test for negro and white students in integrated colleges. *ETS Research Bulletin Series*, 1966(2):i–23, 1966.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pp. 797–806, 2017.

Cortes, C., DeSalvo, G., and Mohri, M. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pp. 67–82. Springer, 2016.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Franc, V. and Prusa, D. On discriminative learning of prediction uncertainty. In *International Conference on Machine Learning*, pp. 1963–1971. PMLR, 2019.

Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. In *Advances in neural information processing systems*, pp. 4878–4887, 2017.

Grari, V., Ruf, B., Lamprier, S., and Detyniecki, M. Fairness-aware neural Réyni minimization for continuous features. *arXiv preprint arXiv:1911.04929*, 2019.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29*, pp. 3315–3323, Barcelona, Spain, December 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Huang, S.-L., Makur, A., Wornell, G. W., and Zheng, L. On universal features for high-dimensional learning and inference. Preprint, 2019. http://allegro.mit.edu/~gww/unifeatures.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 590–597, 2019.

Jones, E., Sagawa, S., Koh, P. W., Kumar, A., and Liang, P. Selective classification can magnify disparities across groups. *arXiv preprint arXiv:2010.14134*, 2020.

Kohavi, R. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pp. 202–207, 1996.

Lee, J., Bu, Y., Sattigeri, P., Panda, R., Wornell, G., Karlinsky, L., and Feris, R. A maximal correlation approach

to imposing fairness in machine learning. *arXiv preprint arXiv:2012.15259*, 2020.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Mary, J., Calauzenes, C., and El Karoui, N. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, pp. 4382–4391, 2019.

Meade, R. Bias in machine learning: How facial recognition models show signs of racism, sexism and ageism. 2019. https://towardsdatascience.com/bias-in-machine-learning-how-facial-recognition-models-show-signs-of-racism-sexism-and-ageism-32549e2c972d.

Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pp. 107–118. PMLR, 2018.

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., and Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pp. 151–159, 2020.

Pessach, D. and Shmueli, E. Algorithmic fairness. *arXiv preprint arXiv:2001.09784*, 2020.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 59–68, 2019.

Turc, I., Chang, M.-W., Lee, K., and Toutanova, K. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*, 2019.

Yildirim, M. Y., Ozer, M., and Davulcu, H. Leveraging uncertainty in deep learning for selective classification. *arXiv preprint arXiv:1905.09509*, 2019.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pp. 962–970. PMLR, 2017.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013a.

Zemel, R., Wu, Y. L., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *Proceedings of the International Conference on Machine Learning*, pp. 325–333, Atlanta, USA, June 2013b.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.