Accounting for Model Uncertainty in Algorithmic Discrimination

Junaid Ali
junaid@mpi-sws.org
Max Planck Institute for Software Systems
Germany

Preethi Lahoti
plahoti@mpi-inf.mpg.de
Max Planck Institute for Informatics
Germany

Krishna P. Gummadi
gummadi@mpi-sws.org
Max Planck Institute for Software Systems
Germany

ABSTRACT

Traditional approaches to ensure group fairness in algorithmic decision making aim to equalize "total" error rates for different subgroups in the population. In contrast, we argue that the fairness approaches should instead focus only on equalizing errors arising due to *model uncertainty* (a.k.a epistemic uncertainty), caused due to lack of knowledge about the best model or due to lack of data. In other words, our proposal calls for ignoring the errors that occur due to uncertainty inherent in the *data*, i.e., aleatoric uncertainty. We draw a connection between *predictive multiplicity* and *model uncertainty* and argue that the techniques from predictive multiplicity could be used to identify errors made due to model uncertainty. We propose scalable convex proxies to come up with classifiers that exhibit predictive multiplicity and empirically show that our methods are comparable in performance and up to four orders of magnitude faster than the current state-of-the-art. We further propose methods to achieve our goal of equalizing group error rates arising due to model uncertainty in algorithmic decision making and demonstrate the effectiveness of these methods using synthetic and real-world datasets.

CCS CONCEPTS

• **Social and professional topics**; • **Computing methodologies** → **Ensemble methods**;

KEYWORDS

algorithmic fairness; classification; model uncertainty; predictive multiplicity

ACM Reference Format:
Junaid Ali, Preethi Lahoti, and Krishna P. Gummadi. 2021. Accounting for Model Uncertainty in Algorithmic Discrimination. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA* . ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3461702.3462630

1 INTRODUCTION

Prediction systems are being used for several socially impactful tasks, e.g., predicting recidivism risk in order to help judges make bail decisions, assessing credit ratings, assessing the risk of defaulting on a loan and predicting the risk of accident for insurance purposes. This development has raised concerns about prediction systems being discriminatory. To address this concern, researchers have proposed a class of group fairness methods, which seek to equalize overall errors across different groups of sensitive attributes such as gender or race [1, 13, 27, 29]. This approach treats all errors as equal. However, not all errors are the same.

It is well-known that errors in prediction models arise out of both epistemic (model) uncertainty and aleatoric (inherent) uncertainty [5, 14, 20]. Equalizing total error could lead to unjustifiably wrong decisions for some datapoints. Consider Figure 1, where a traditional fair classifier that equalizes total errors including the irreducible ones that arise due to aleatoric uncertainty. This results in many datapoints getting a negative outcome even though they clearly belong to the positive cluster. These errors are particularly consequential in socially impactful applications.

In this paper, we argue to distinguish between the errors caused by different types of uncertainty. Specifically, we introduce the notions of *aleatoric errors* and *epistemic errors*. We refer to the errors that occur only due to model or epistemic uncertainty as *epistemic errors* and the ones that occur due to aleatoric uncertainty, we call the *aleatoric errors*. Figure 1 shows an example of both types of errors. The errors made by the classifiers C_1 and C_2 that are highlighted by the region **A** are due to the noise in the data, as these wrongly predicted datapoints are surrounded by predominantly the other class label, i.e., ground truth positive or ground truth negative datapoints. We refer to these types of errors as *aleatoric errors*. While the errors in the region marked by **E** are due to model uncertainty as one could resolve this uncertainty by gathering more data or by choosing a more complex model. These types of errors are *epistemic errors*. Our proposal is to *ignore* the aleatoric errors which are likely to be irreducible due to inherent uncertainty in the data or the prediction task at hand and we argue to *only* equalize the epistemic errors, i.e., the ones that occur due to methodological limitations.

In order to identify the epistemic errors that are caused by model uncertainty, we leverage the work on predictive multiplicity by Marx et al. [21]. *Predictive multiplicity* refers to the scenario where multiple predictive models have similar predictive performance (e.g., similarly accurate) but assign contradictory predictions on a subset of the datapoints, which characterize the *ambiguous regions*. We draw a connection between predictive multiplicity and *model uncertainty*.

AIES '21, May 19–21, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8473-5/21/05.
https://doi.org/10.1145/3461702.3462630

Model uncertainty is defined as the level of spread or 'disagreement' in the decisions of an ensemble sampled from the posterior [20]. We use predictive multiplicity to identify model uncertainty, i.e., we argue that the disagreement in equally well performing models signals uncertainty in the model parameters. Specifically, we argue that if the classifiers exhibiting predictive multiplicity are chosen from a complex enough hypothesis class, then the regions in the feature space with high model uncertainty that are likely to have the epistemic errors would coincide with the ambiguous regions produced by predictive multiplicity. Therefore, our proposal of equalizing only the epistemic errors translates into equalizing errors in the ambiguous regions, while ignoring the ones in the unambiguous regions.

One of the *key properties* of our proposal is that people whose outcomes are affected by our fairness requirements are the people whose outcomes are ambiguous or uncertain in the first place. Put differently, we do not alter the outcomes of people with unambiguously positive or negative outcomes. In contrast, current methods for achieving equal error rates might alter outcomes for people with unambiguous outcomes as well, creating a difficult accuracy-fairness tradeoff dilemma. We believe that our proposal would be easier to justify in many practical scenarios.

Key technical contributions of our approach are (a) designing efficient and scalable methods for identifying ambiguous regions, and (b) designing mechanisms for equalizing group error rates in the ambiguous regions. In order to solve the first challenge, we propose *convex proxies* to find models that exhibit predictive multiplicity. For the second challenge, our key insight is *to reuse the highly accurate models trained to identify the ambiguous regions* in the first place. Specifically, given the set of classifiers identifying ambiguous regions, we propose to *stochastically pick a classifier* from this set when making a decision. The probabilities of picking the classifiers are chosen in a way that equalizes group error rates in the ambiguous regions in expectation. An additional benefit of our approach compared to the traditional way of making a deterministic decision is that we account for model uncertainty by introducing stochasticity in our predictions, and thus many datapoints in the *ambiguous region* have a non-zero probability of receiving a favorable outcome. As there is some chance of getting a favorable outcome for most datapoints affected by our fairness notion, it would make our proposal more desirable than the traditional approach of assigning decisions deterministically.

Contributions and Outline:

- **Conceptual contribution:** We argue that uncertainty in prediction should be accounted for when designing fairness approaches. To this end, we propose to *only* equalize errors occurring due to model uncertainty, i.e., the epistemic errors.
- **Technical contributions:** i) We propose tractable scalable convex proxies to identify *ambiguous regions*. That is, for a given dataset $\mathcal{D}$, we identify a set of linear or nonlinear classifiers that are equally accurate, but they conflict in their predictions for a subset of datapoints (see Section 3.1). ii) We also formalize our proposal to only equalize the epistemic errors and present a fairness approach that equalizes group errors in the *ambiguous regions* (see Section 3.2).

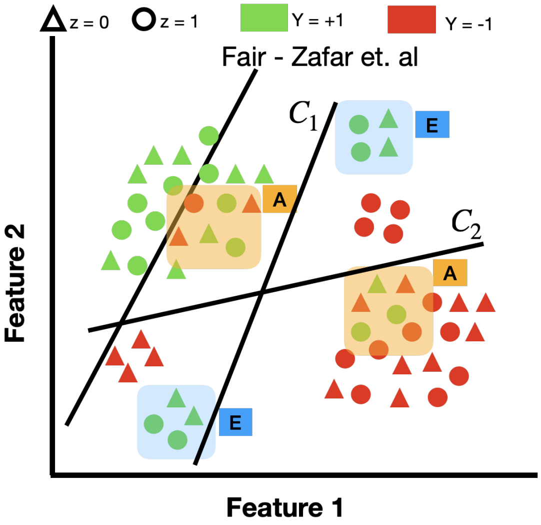

Figure 1: Illustrative example: Consider a binary classification task with two features and a sensitive feature (z) represented by the shape of the points, i.e., circles and triangles. Green and red colors represent ground truth positive and negative labels, respectively. Classifiers C_1 and C_2 are equally accurate classifiers achieving 80% accuracy. The difference between false positives of triangles and circles for C_1 is 32% and −25% with C_2. However, these two classifiers disagree on their decision on 29% of the data, i.e., which lies in the ambiguous region between the two classifiers. The errors made by these classifiers in the ambiguous regions marked by E are *epistemic errors*. While the errors highlighted by the region A are *aleatoric errors*. If we were to pick one of these classifiers as the final decision boundary it would be unfair to the points receiving a favorable decision with the other classifier. On the other hand, a fair classifier equalizing false positive rates, using [29], gives an accuracy of only 68%. However, as it does not disregard the aleatoric uncertainty it changes the decisions of several points that clearly belong to the positive cluster.

- **Empirical contributions:** i) Our experimental results show that our proposed scalable convex proxies to identify regions with predictive multiplicity are comparable in performance and up to four orders of magnitude faster than the current state-of-the-art (see Section 4.4, Table 2). ii) Our experimental results on a synthetic and two real-world datasets show that our fairness methods improve fairness in the ambiguous regions while achieving comparable accuracy to the best classifier (see Sections 4.4 and 4.5).

2 PRELIMINARIES AND BACKGROUND

In this section, we present the necessary background on binary classification and predictive multiplicity.

2.1 Binary Classification

Given a training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, the goal of a binary classifier is to learn a function $\boldsymbol{\phi} : \mathbb{R}^d \rightarrow \{-1, 1\}$ between the feature vectors $\boldsymbol{x} \in \mathbb{R}^d$ and the class labels $y \in \{-1, 1\}$. In order to learn this function one has to solve $\boldsymbol{\phi}^* = \operatorname{argmin}_{\boldsymbol{\phi}} \boldsymbol{R}_{\mathcal{D}}(\boldsymbol{\phi}) : \boldsymbol{R}_{\mathcal{D}}(\boldsymbol{\phi}) = \frac{1}{N} \sum_{\boldsymbol{x}_i, y_i} \mathbb{1}[\boldsymbol{\phi}(\boldsymbol{x}_i) \neq y_i]$. However, this function is non-convex in $\boldsymbol{\phi}$ and worse, it is intractable, which makes it especially difficult to solve for large datasets. In the rest of the text we drop the subscript, $\mathcal{D}$, for brevity. To efficiently solve the problem, it is a standard practice to use a convex proxy. One minimizes a given convex loss $L(\boldsymbol{\theta})$ over $\mathcal{D}$, *i.e.*, $\boldsymbol{\theta}^* = \operatorname{argmin}_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$, in order to find $\boldsymbol{\theta}^*$ for convex decision-boundary-based classifiers like linear/non-linear SVM and logistic regression, where $\boldsymbol{\theta} \in \mathbb{R}^d$. Then, for a given (potentially unseen) feature vector $\boldsymbol{x}$, one predicts the class label $\hat{y} = 1$ if $d_{\boldsymbol{\theta}^*}(\boldsymbol{x}) \geq 0$ and $\hat{y} = -1$ otherwise, where $d_{\boldsymbol{\theta}^*}(\boldsymbol{x})$ denotes the signed distance from $\boldsymbol{x}$ to the decision boundary. For convenience, we define $\boldsymbol{\theta}^*(\boldsymbol{x}) = 1$ if $d_{\boldsymbol{\theta}^*}(\boldsymbol{x}) \geq 0$ and $\boldsymbol{\theta}^*(\boldsymbol{x}) = -1$ if $d_{\boldsymbol{\theta}^*}(\boldsymbol{x}) < 0$.

In the rest of the paper, we consider $\boldsymbol{\theta}_{best}$ to be the most accurate classifier yielded by minimizing logistic regression loss with L2 regularizer, where weights of the regularizer were picked based on the performance on the validation set. Similarly, we consider $\boldsymbol{\phi}_{best}$ to be the best classifier using 0-1 loss ($\boldsymbol{R}_{\mathcal{D}}$), selected using a validation set.

2.2 Background on Predictive Multiplicity

In this section, we formally introduce the notion of predictive multiplicity and discuss the existing measures and mechanisms to compute predictive multiplicity.

Predictive multiplicity. A prediction problem exhibits predictive multiplicity if one can find a classifier $\boldsymbol{\phi}$ for a given small value $\boldsymbol{\epsilon}$ such that $\boldsymbol{R}(\boldsymbol{\phi}) - \boldsymbol{R}(\boldsymbol{\phi}_{best}) <= \boldsymbol{\epsilon}$, and there exists at least one datapoint with feature vector $\boldsymbol{x}_i$ such that $\boldsymbol{\phi}(\boldsymbol{x}_i) \neq \boldsymbol{\phi}_{best}(\boldsymbol{x}_i)$ [21]. The definition for classifiers trained with proxy loses is similar. One could consider $\boldsymbol{\epsilon}$ to be 0 but in practice a classifier that is slightly less accurate on the training data might be equally or even more accurate on the test data.

Predictive multiplicity is defined for a set of two or more classifiers, referred to as the $\boldsymbol{\epsilon}$-level set. Given the most accurate classifier $\boldsymbol{\phi}_{best}$, the $\boldsymbol{\epsilon}$-level set of $\boldsymbol{\phi}_{best}$ is a set of classifiers which have an accuracy only up to $\boldsymbol{\epsilon}$ lower than $\boldsymbol{\phi}_{best}$. Formally, over the dataset $\mathcal{D}$, $\mathbb{C}_{\boldsymbol{\epsilon}, \boldsymbol{\phi}_{best}} = \{\boldsymbol{\phi} : \boldsymbol{R}(\boldsymbol{\phi}) - \boldsymbol{R}(\boldsymbol{\phi}_{best}) \leq \boldsymbol{\epsilon}\}$.

Measures of predictive multiplicity. Marx et al. [21] propose two measures for predictive multiplicity for a given set of classifiers, namely *Discrepancy* and *Ambiguity*.

For a given set of classifiers, *Discrepancy* is defined as the maximum fraction of the datapoints on which any classifier in the set disagrees on the outcomes with the most accurate classifier. Formally, given $\mathbb{C}_{\boldsymbol{\epsilon}, \boldsymbol{\phi}_{best}}$ and dataset $\mathcal{D}$,

$$\delta_{\boldsymbol{\epsilon}}(\boldsymbol{\phi}) = \max_{\boldsymbol{\phi} \in \mathbb{C}_{\boldsymbol{\epsilon}}} \frac{1}{n} \sum_{\boldsymbol{x}_i \in \mathcal{D}} \mathbb{1}[\boldsymbol{\phi}(\boldsymbol{x}_i) \neq \boldsymbol{\phi}_{best}(\boldsymbol{x}_i)], \tag{1}$$

i.e., discrepancy is the maximum fraction of conflicting decisions yielded by any classifier in $\mathbb{C}_{\boldsymbol{\epsilon}, \boldsymbol{\phi}_{best}}$ compared to $\boldsymbol{\phi}_{best}$.

Ambiguity of a set of classifiers for a prediction task is defined as the fraction of datapoints given a different decision than the best classifier. Formally, given set $\mathbb{C}_{\boldsymbol{\epsilon}, \boldsymbol{\phi}_{best}}$ and dataset $\mathcal{D}$,

$$\alpha_{\boldsymbol{\epsilon}}(\boldsymbol{\phi}) = \frac{1}{n} \sum_{\boldsymbol{x}_i} \max_{\boldsymbol{\phi} \in \mathbb{C}_{\boldsymbol{\epsilon}, \boldsymbol{\phi}_{best}}} \mathbb{1}[\boldsymbol{\phi}(\boldsymbol{x}_i) \neq \boldsymbol{\phi}_{best}(\boldsymbol{x}_i)], \tag{2}$$

where $\max_{\boldsymbol{\phi} \in \mathbb{C}_{\boldsymbol{\epsilon}, \boldsymbol{\phi}_{best}}} \mathbb{1}[\boldsymbol{\phi}(\boldsymbol{x}_i) \neq \boldsymbol{\phi}_{best}(\boldsymbol{x}_i)]$ is 1 if there exists at least one classifier in $\mathbb{C}_{\boldsymbol{\epsilon}, \boldsymbol{\phi}_{best}}$ which gives a datapoint with features $\boldsymbol{x}_i$ a different outcome than $\boldsymbol{\phi}_{best}$, otherwise it is 0. Hence, ambiguity is the fraction of datapoints on which any classifiers in $\mathbb{C}_{\boldsymbol{\epsilon}, \boldsymbol{\phi}_{best}}$ disagrees on the outcome with ϕ_{best}.

Methods to identify predictive multiplicity. Inspired by the measures discrepancy and ambiguity, Marx et al. [21] propose two methods that maximize these measures in order to find the classifiers that exhibit maximum predictive multiplicity for the given allowance of accuracy reduction. This would indicate the extent of predictive multiplicity for the prediction task at hand.

Exact discrepancy maximization (Dsc-Exact). Given a value of $\boldsymbol{\epsilon}$, the authors propose to train classifiers that minimize the agreement to $\boldsymbol{\phi}_{best}$ under the constraint that its accuracy is only up to $\boldsymbol{\epsilon}$ lower than $\boldsymbol{\phi}_{best}$, i.e.,

$$\begin{aligned} &\underbrace{\underset{\boldsymbol{\phi}}{\text{minimize}} \quad \sum_{\boldsymbol{x}_i} \mathbb{1}[\boldsymbol{\phi}(\boldsymbol{x}_i) = \boldsymbol{\phi}_{best}]}_{\text{maximize discrepancy}} \\ &\text{subject to} \quad \underbrace{\boldsymbol{R}(\boldsymbol{\phi}) \leq \boldsymbol{R}(\boldsymbol{\phi}_{best}) + \boldsymbol{\eta}}_{\text{bound accuracy reduction}} \end{aligned} \tag{P1}$$

where $\boldsymbol{\eta} \in (0, \boldsymbol{\epsilon})$. One can obtain a set $\mathbb{C}_{\boldsymbol{\epsilon}, \boldsymbol{\phi}_{best}}$ by solving the above formulation for several $\boldsymbol{\eta}$ values.

Exact ambiguity maximization (Amb-Exact). In order to find the classifiers that maximize the ambiguity measure for a given threshold of accuracy reduction, Marx et al. [21] propose to train a classifier for each datatpoint in the training data that gives the datapoint a different decision than the most accurate classifier. Then, they pick the classifiers whose accuracy lies within the threshold of the allowed accuracy reduction. Specifically, they propose to train classifiers that change their decisions compared to $\boldsymbol{\phi}_{best}$ for individual datapoints while minimizing 0-1 loss, i.e.,

$$\underbrace{\underset{\boldsymbol{\phi}}{\text{minimize}}\ \boldsymbol{R}(\boldsymbol{\phi})}_{\text{maximize accuracy}} \quad \text{subject to} \quad \underbrace{\boldsymbol{\phi}(\boldsymbol{x}_i) \neq \boldsymbol{\phi}_{best}(\boldsymbol{x}_i)}_{\text{change decision of } \boldsymbol{x}_i \text{ w.r.t } \boldsymbol{\phi}_{best}} \quad \forall \boldsymbol{x}_i. \tag{P2}$$

Then, one can select $\mathbb{C}_{\boldsymbol{\epsilon}, \boldsymbol{\phi}_{best}}$ by pruning the set of classifiers resulting from the solution of the problem above, i.e., by selecting classifiers which are only $\boldsymbol{\epsilon}$ lower in accuracy than $\boldsymbol{\phi}_{best}$.

To solve both Problems P1 and P2, Marx et al. [21] propose mixed integer programming formulations. However, these formulations i) work only for linear classifiers and ii) have slow performance as these are exact, intractable and non-convex.

3 PROPOSED APPROACH

In this section, we aim to answer the question: *What is a fair model under model uncertainty*?

We characterize model uncertainty using predictive multiplicity. Given a set of classifiers $\mathbb{C}_{\boldsymbol{\epsilon}, \boldsymbol{\theta}_{best}}$ that exhibit predictive multiplicity,

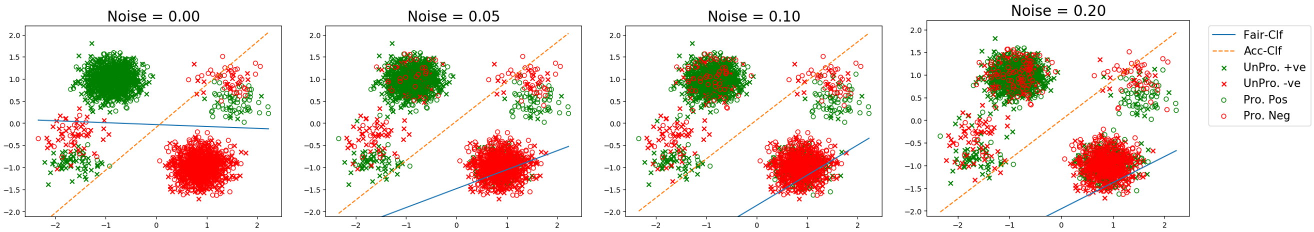

Figure 2: [Synthetic dataset] Figure demonstrates that state of the art fairness methods are effected by label noise.

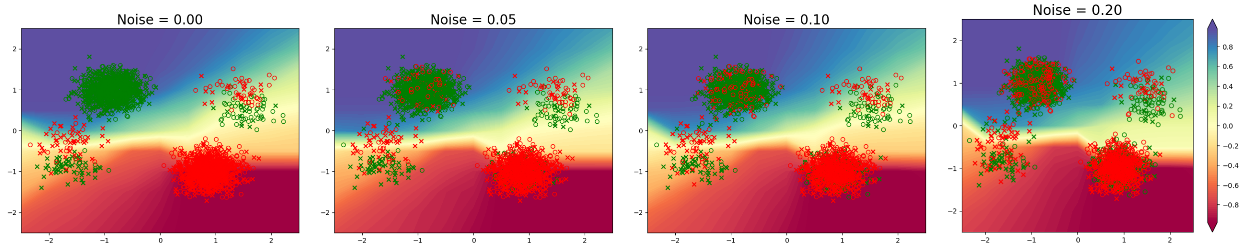

Figure 3: [Synthetic dataset] Figure shows the expected class while equalizing FPRs using the classifiers solving P4. It demonstrates that our method is stable under label noise, as it consistently identifies same regions as ambiguous for different levels of noise values.

we consider x_i to have an ambiguous decision if *any* of the classifiers in $\mathbb{C}_{\epsilon,\theta_{best}}$ gives it a conflicting decision compared to any other classifier. Formally a set of ambiguous points are defined as:

$$\mathcal{A} := \{x_i : \theta_j(x_i) \neq \theta_k(x_i) \,\forall\, \theta_j, \theta_k \in \mathbb{C}_{\epsilon,\theta_{best}}\}.$$

These points characterize the ambiguous region. By choosing a single model from $\mathbb{C}_{\epsilon,\theta_{best}}$ as the final model we might be unfair to some group in the ambiguous region. Our proposal of only equalizing the epistemic errors boils down to equalizing group error rates in the ambiguous region $\mathcal{A}$.

The *key assumption* we make is that the hypothesis class for the classifiers, $\mathbb{C}_{\epsilon,\theta_{best}}$, exhibiting predictive multiplicity is sufficiently complex, i.e., if the data is nonlinearly separable the hypothesis class should include nonlinear classifiers. Under this assumption, all the errors in the the unambiguous region, i.e., where all the classifiers in the set $\mathbb{C}_{\epsilon,\theta_{best}}$ agree in their decisions, would *only* be due aleatoric uncertainty. The argument is as follows: Given the classifiers in set $\mathbb{C}_{\epsilon,\theta_{best}}$ are picked from a sufficiently complex hypothesis class for the given data. Under this assumption, if all the classifiers agree in their prediction for a subset of the datapoints, then the resulting errors for these datapoints could only be due to inherent stochasticity of the prediction task or random noise, i.e., aleatoric errors. On the other hand, the ambiguous region, $\mathcal{A}$, would identify regions with high model uncertainty. The intuition is as follows: Given that the classifiers for set $\mathbb{C}_{\epsilon,\theta_{best}}$ are chosen from a sufficiently complex hypothesis class. Under this assumption, if these equally accurate classifiers disagree on some datapoints this would include all the datapoints whose decisions are uncertain due to lack of data. This implies that *all* the epistemic errors will lie in the ambiguous region. The ambiguous region could also have random noise hence causing some aleatoric errors. The results using the Synthetic dataset in Section 4.4 confirm our hypotheses.

Next, we present our proposals for identifying the ambiguous region using scalable convex methods. Then, we discuss our methods for equalizing groups errors in the ambiguous region $\mathcal{A}$.

3.1 Scalable Methods for Predictive Multiplicity

In this section, we propose two convex methods to find the ambiguous region $\mathcal{A}$.

Approximate Discrepancy maximization (Dsc-Approx). We propose the following convex and tractable proxy constraint that bounds similarity between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{best}$, akin to the objective in Problem P1 that maximizes discrepancy:

$$\frac{1}{N}\sum_{x} \max(0, d_{\boldsymbol{\theta}(x)} d_{\boldsymbol{\theta}_{best}(x)}) \leq \gamma, \tag{3}$$

where $d_{\boldsymbol{\theta}(x)}$ represents the distance of the datapoint with feature vector $\boldsymbol{x}$ from the decision boundary of $\boldsymbol{\theta}$. $\max(0,\cdot)$ represents the agreement of decisions between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{best}$. Specifically, if the decision for a subject with feature vector $\boldsymbol{x}$ stays the same under $\boldsymbol{\theta}$ compared to $\boldsymbol{\theta}_{best}$, only then does the term $\max(0,\cdot)$ produce a non-zero number. Thus, by bounding the left hand side we are limiting the average allowed distance of the datapoints which have the same decisions under $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{best}$. Making this bound tighter would preferably admit $\boldsymbol{\theta}$ whose decisions are different on some of the datapoints than $\boldsymbol{\theta}_{best}$, as those datapoints contribute 0 to the sum on the left hand side. This implies that one can control the number of decisions allowed to be the same between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{best}$

by changing the value of $\gamma \in \mathbb{R}+$. For example, $\gamma = +\infty$ would yield $\boldsymbol{\theta} = \boldsymbol{\theta}_{best}$ meaning that all the decisions between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{best}$ are the same, i.e., $\boldsymbol{\theta}$ would yield a discrepancy of 0 compared to $\boldsymbol{\theta}_{best}$. Similarly, for $\gamma = 0$ one aims to learn $\boldsymbol{\theta}$ whose decisions are different on all datapoints than to $\boldsymbol{\theta}_{best}$, i.e. a classifier yielding maximum discrepancy compared to $\boldsymbol{\theta}_{best}$. The value of γ also controls the reduction in accuracy under $\boldsymbol{\theta}$ compared to $\boldsymbol{\theta}_{best}$.

For linear boundary-based classifiers (logistic regression, linear SVM), $d_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^T \boldsymbol{x}$. For nonlinear SVM, one can write $d_{\beta}(\boldsymbol{x}) = \sum_{i=1}^{N} \beta_i y_i k(\boldsymbol{x}_i, \boldsymbol{x})$ for the optimization variables β and a positive semidefinite kernel function $k(.,.)$. Hence, in both linear and nonlinear cases the constraint stays convex since the distance from the decision boundary is linear with respect to the optimization variables.

One can write a convex and tractable version of Problem P1 using the logistic regression loss as follows:

$$\begin{aligned} &\underbrace{\underset{\boldsymbol{\theta}}{\text{minimize}} \quad -\frac{1}{N}\sum_{\boldsymbol{x}_i, y_i} p(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})}_{\text{maximize accuracy}} \qquad \text{(P3)} \\ &\text{subject to} \quad \underbrace{\frac{1}{N}\sum_{\boldsymbol{x}_i} \max(0, d_{\boldsymbol{\theta}(\boldsymbol{x}_i)} d_{\boldsymbol{\theta}_{best}(\boldsymbol{x}_i)}) \leq \gamma}_{\text{enforce discrepancy}} \end{aligned}$$

where $p(y = 1|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{1+\exp(-\boldsymbol{\theta}^T \boldsymbol{x})}$.

One can learn an appropriate γ value using a validation set, for a given $\boldsymbol{\eta}$ in P1. We construct $\mathbb{C}_{\boldsymbol{\epsilon}, \boldsymbol{\theta}_{best}}$ by training classifiers with varying values of γ and then picking the ones whose accuracy is only $\boldsymbol{\epsilon}$ lower than $\boldsymbol{\theta}_{best}$.

Approximate ambiguity maximization (Amb-Approx). We propose the following convex and tractable constraints equivalent to the constraint in Problem P2.

$$\begin{aligned} d_{\boldsymbol{\theta}(\boldsymbol{x}_i)} &< 0 \text{ if } d_{\boldsymbol{\theta}_{best}(\boldsymbol{x}_i)} \geq 0 \quad \forall \boldsymbol{x}_i \\ d_{\boldsymbol{\theta}(\boldsymbol{x}_i)} &\geq 0 \text{ if } d_{\boldsymbol{\theta}_{best}(\boldsymbol{x}_i)} < 0 \quad \forall \boldsymbol{x}_i, \end{aligned} \qquad (4)$$

where $d_{\boldsymbol{\theta}}$ is the distance from decisions boundary of $\boldsymbol{\theta}$. The constraints above require $\boldsymbol{\theta}$ to make a different decision than $\boldsymbol{\theta}_{best}$ on the datapoint $\boldsymbol{x}_i$. The constraints stay convex for both linear and nonlinear boundary based classifiers because one can write the distance from the decision boundary as a linear function of the optimization parameter in both cases. One can write a convex and scalable version of Problem P2 as follows:

$$\begin{aligned} &\underbrace{\underset{\boldsymbol{\theta}}{\text{minimize}} \quad -\frac{1}{N}\sum_{\boldsymbol{x}_i, y_i} p(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})}_{\text{maximize accuracy}} \qquad \text{(P4)} \\ &\text{subject to} \quad \underbrace{\begin{aligned} d_{\boldsymbol{\theta}(\boldsymbol{x}_i)} &< 0 \text{ if } d_{\boldsymbol{\theta}_{best}(\boldsymbol{x}_i)} \geq 0 \quad \forall \boldsymbol{x}_i \\ d_{\boldsymbol{\theta}(\boldsymbol{x}_i)} &\geq 0 \text{ if } d_{\boldsymbol{\theta}_{best}(\boldsymbol{x}_i)} < 0 \quad \forall \boldsymbol{x}_i, \end{aligned}}_{\text{change decision of } \boldsymbol{x}_i \text{ w.r.t } \boldsymbol{\theta}_{best}} \end{aligned}$$

where $p(y = 1|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{1+\exp(-\boldsymbol{\theta}^T \boldsymbol{x})}$. We pick $\mathbb{C}_{\boldsymbol{\epsilon}, \boldsymbol{\theta}_{best}}$ by training a set of classifiers which assign conflicting decisions to all the datapoints in the training set. Then, we pick the classifiers which are only $\boldsymbol{\epsilon}$ lower in accuracy than $\boldsymbol{\theta}_{best}$.

3.2 Leveraging Predictive Multiplicity towards Fairness under Model Uncertainty

In this section, we propose to learn a meta classifier in order to equalize group errors arising due to model uncertainty.

In order to do that, our key insight is to use the highly accurate classifiers that we trained to identify the ambiguous regions in the first place. Specifically, given the validation set of datapoints and $\mathbb{C}_{\boldsymbol{\epsilon}, \boldsymbol{\theta}_{best}}$, picked by solving Dsc-Approx, P3, or Amb-Approx, P4, we first identify the points with ambiguous decisions. We then construct a meta classifier by picking the classifiers stochastically from the set $\mathbb{C}_{\boldsymbol{\epsilon}, \boldsymbol{\theta}_{best}}$. The probabilities for picking these classifiers are chosen in a way that aims to equalize group error rates on the ambiguous datapoints among different groups of a sensitive feature such as race or gender. For a binary valued sensitive feature $z = \{0, 1\}$, we propose

$$\begin{aligned} &\underset{w}{\text{minimize}} \quad |\sum_{\theta \in \mathbb{C}_\epsilon} w_\theta \cdot \underbrace{(Err_{z=1}(\theta) - Err_{z=0}(\theta))}_{\text{FPR/FNR difference}}| \qquad \text{(P5)} \\ &\text{subject to} \quad 0 \leq w_\theta \leq 1 \quad \text{and} \quad \sum_{\theta} w_\theta = 1, \end{aligned}$$

where $Err_{z=0}(\theta)$ and $Err_{z=1}(\theta)$ are false positive rates (FPR) or false negative rates (FNR) for group 0 and 1 of the sensitive feature *in the ambiguous region, $\mathcal{A}$*. As the set of classifiers is predetermined, the error rates can be precomputed. Hence, the problem is convex and efficiently solvable, as the objective function is a linear function of optimization variable w.

The intuition is that the difference of the errors rates between the two groups, i.e., $Err_{z=1}(\theta) - Err_{z=0}(\theta)$, might be positive for some of the classifiers in $\mathbb{C}_{\boldsymbol{\epsilon}, \boldsymbol{\theta}_{best}}$ and it might be negative for the others. We can then assign the probabilities w_θ to these classifiers in a way such that they cancel each others biases and the expected unfairness is minimized. Our experimental results on the real-world and synthetic datasets confirm our intuition (Tables 1, 3, 4).

In the case of a non-binary valued sensitive feature, one can replace the error rate difference between two groups with pair-wise differences among all the groups. We learn the probability mass function w using the validation datapoints, and when classifying the unseen test datapoints we use w to pick the classifiers from $\mathbb{C}_{\boldsymbol{\epsilon}, \boldsymbol{\theta}_{best}}$.

4 EXPERIMENTS

In this section, we demonstrate the effectiveness of our methods using synthetic and real-word datasets. Specifically, we answer the following evaluation questions:
– **Q1.** How effective and fast are our methods in identifying the ambiguous regions?
– **Q2.** What is the fairness and accuracy trade-off of our methods?
– **Q3.** Are our methods robust to noisy data?

4.1 Datasets

We use a *Synthetic dataset* because i) we could easily alter the size of the datasets, which is useful as Dsc-Exact and Amb-Exact have slow performance on larger complex datasets, especially with continuous valued features; ii) we could provide intuition for the type of ambiguous regions identified by our methods; iii) we could introduce noise in the data and check the robustness of our methods vs the existing methods. The data comprises 10000 datapoints and 2 features and a binary valued sensitive feature, z. The data is sampled from the following Gaussian distributions:

$$\mathcal{N}_1([-35; 65], [60, 1; 1, 120]),\quad \mathcal{N}_2([15; -25], [60, 1; 1, 120]),$$
$$\mathcal{N}_3([30; 65], [70, 1; 1, 100]),\quad \mathcal{N}_4([35; 40], [70, 1; 1, 100]),$$
$$\mathcal{N}_5([-55; 5], [70, 1; 1, 100]) \text{ and } \mathcal{N}_6([-55; -20], [70, 1; 1, 100])$$

From $\mathcal{N}_1$, 4500 points were sampled. Amongst these, 95% of which were labeled ground truth positive and 65% of these points were uniformly at random assigned to the non-protected class of the sensitive feature, i.e, $z = 0$. A total of 4500 points were sampled from $\mathcal{N}_2$, 95% of which are ground truth negative points and 65% of these points were uniformly at random assigned to the protected class of the sensitive feature, i.e., $z = 0$. Finally, 250 points were sampled from $\mathcal{N}_3$ and $\mathcal{N}_5$ each, with ground truth negative labels, and 250 points were sampled from $\mathcal{N}_4$ and $\mathcal{N}_6$ each and were assigned ground truth positive labels. 80% of the points sampled from $\mathcal{N}_3$ and $\mathcal{N}_4$ and 20% of the points sampled from $\mathcal{N}_5$ and $\mathcal{N}_6$, were uniformly at randomly assigned $z = 1$. After sampling these points they were normalized to have a unit mean and a unit variance. A visual representation is shown in Figure 2. We flipped the class label of a fraction of datapoints which induced aleatoric errors through out the data. However, model uncertainty only exists in the sparse clusters shown in Figure 2 as that could be reduced by gathering more data. Our hope is that predictive multiplicity would be able to identify regions with predominantly model uncertainty, i.e., the sparse clusters as the ambiguous regions for different levels of aleatoric uncertainty. We also experimented with other variations of the parameters and got similar results.

We processed the ProPublica *COMPAS dataset* [19] similar to Zafar et al. [28], which resulted in 5, 287 subjects and 7 features. Given these features we have to predict whether a criminal defendant would recidivate within two years (positive class) or not (negative class). We consider race, with values African-Americans, $z = 0$, and white, $z = 1$, to be a sensitive feature in this dataset.

The NYPD *SQF dataset* comprises features of pedestrians, such as race, gender, height *etc.* and the goal is to predict whether (negative class) or not (positive class) a weapon was discovered on inspection. We use race as a sensitive feature, z, in our experiments, with African-Americans ($z = 0$) and white ($z = 1$) as two values of this feature. After processing the data similar to Zafar et al. [28] the dataset consists of 5, 832 subjects and 22 features.

4.2 Experimental Setup

The datasets were split into 50% training, 25% validation and 25% test datapoints. Training data was used to train the classifiers, validation data for tuning hyper parameters and test data to report the results. The CVXPy library [7] was used to solve all the formulations. We show results using linear classifiers, as decisions made by the linear classifiers are relatively easier to explain, which is an import goal for applications with social significance such as recidivism risk prediction. Additionally, data are likely to be linearly separable in higher dimensions. We show some results using nonlinear boundaries with our methods in the appendix.

Table 1: [Synthetic dataset] Signed differences in FPR/FNR

	Unfairness			Accuracy
	total	unamb	amb	
Acc.	-0.13/-0.14	0.05/-0.06	0.46/-0.45	0.89
Fair	0.03/-0.02	0.05/-0.06	-0.14/0.29	0.77/0.89
Uni-P3	0.04/-0.04	0.05/-0.06	-0.22/0.20	0.89 / 0.89
Our-P3	0.07/-0.07	0.05/-0.06	**0.0/-0.01**	0.89/0.89
With P4				
Acc.	0.13/-0.14	0.06/-0.07	0.30/-0.35	0.89
Fair	0.03/-0.02	0.05/-0.07	-0.06/0.18	0.77/0.89
Uni-P4	0.10/-0.10	0.06/-0.07	0.16/-0.16	0.88 / 0.88
Our-P4	0.06/-0.07	0.06/-0.07	**0.01/-0.03**	0.88/ 0.88

This table demonstrates that our method is effective in removing unfairness at a very small cost of decrease in the accuracy. Please refer to Section 4.4

Selecting $\mathbb{C}_{\epsilon,\theta_{best}}$. We generate $\mathbb{C}_{\epsilon,\theta_{best}}$ by solving Dsc-Approx, given by Problem P3, for a range of γ values or Amb-Approx, given by Problem P4, for each training datapoint. Then, we use the validation data to prune the resulting classifiers which lie within a given ϵ threshold of the most accurate classifier. The results are averaged over 5 runs of these steps using different seed values to initialize the data-split and the solver. For Dsc-Approx, we pick the $\mathbb{C}_{\epsilon}$ from the aggregated solutions of all the seeds and present the averaged statistics over all the seeds.

We assume that ϵ is chosen by the experts for the prediction task at hand. We present results for $\epsilon = 0.02$ for the synthetic dataset, and $\epsilon = 0.01$ for real-world datasets. We experimented with several values of ϵ and obtained similar results.

4.3 Benchmarks and Metrics

In this section, we discuss the benchmarks and metrics we used to evaluate our proposals.

Ambiguous regions computation benchmarks. In order to demonstrate the efficiency of our methods to identify the ambiguous regions using Dsc-Approx and Amb-Approx, we compare with Dsc-Exact and Amb-Exact. We solved the Dsc-Exact and Amb-Exact problems using the CPLEX library, with the code provided by the authors [21].

Metrics for evaluating ambiguous regions computation. Since the best classifiers for non-scalable and our scalable methods, i.e., $\boldsymbol{\phi}_{best}$ and $\boldsymbol{\theta}_{best}$, are different, we report the ambiguity $\hat{\alpha}$ and discrepancy $\hat{\delta}$ between any two classifiers in $\mathbb{C}_{\epsilon}$, for the respective methods. They are formally defined as follows:

$$\hat{\delta}_{\boldsymbol{\epsilon}}(\boldsymbol{\phi}) = \max_{\boldsymbol{\phi},\hat{\boldsymbol{\phi}}\in\mathbb{C}_{\boldsymbol{\epsilon}}} \frac{1}{n}\sum_{\boldsymbol{x}_i} \mathbb{1}[\boldsymbol{\phi}(\boldsymbol{x}_i) \neq \hat{\boldsymbol{\phi}}(\boldsymbol{x}_i)] \tag{5}$$

Table 2: Comparison identifying ambiguous regions

ϵ	P1		P2		P3		P4	
-	$\hat{\delta}$	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\alpha}$
0.03	0.15	0.16	**0.18**	**0.28**	0.14	0.16	**0.18**	0.26
0.05	0.17	0.19	0.22	**0.38**	0.16	0.17	**0.23**	0.36
0.09	0.22	0.24	**0.32**	**0.56**	0.2	0.20	**0.32**	0.51

Training Time

Time	P1	P2	P3	P4
mins	510	19227	5	5

The table above shows maximum discrepancy and ambiguity between any two classifiers in the $\mathbb{C}_{\epsilon,\psi:\psi\in\{\phi_{best},\theta_{best}\}}$. The bottom table shows the time it took to compute the ambiguous regions with each method. It shows that our methods, given by P3 and P4, achieve comparable performance compared to P1 and P2 and they are upto four orders of magnitude faster. Please refer to Section 4.4

Table 3: [COMPAS] Signed differences in FPR/FNR

	Unfairness			Accuracy
	total	unamb	amb	
Acc.	-0.19/0.33	-0.23/0.41	**0.08**/-0.20	0.66
Fair	0.02/0.03	-0.09/0.18	0.83/-0.92	0.66/0.65
Uni-P3	-0.20/0.35	-0.23/0.41	**-0.08**/**-0.004**	0.66 /0.66
Our-P3	-0.20/0.35	-0.23/0.41	**-0.08**/-0.02	0.66/0.66
With P4				
Acc.	-0.19/0.33	-0.24/0.54	-0.11/0.15	0.66
Fair	0.02/0.03	-0.24/0.54	0.34/0.-0.42	0.66/0.65
Uni-P4	-0.19/0.34	-0.24/0.54	-0.11/0.15	0.66/ 0.66
Our-P4	-0.14/0.26	-0.24/0.54	**-0.01**/**0.03**	0.66/ 0.66

This table demonstrates that our methods are effective in removing unfairness in the ambiguous regions at no expense of accuracy. Please refer to Section 4.5

$$\hat{\alpha}_{\epsilon}(\phi) = \frac{1}{n}\sum_{x_i}\max_{\phi,\hat{\phi}\in\mathbb{C}_{\epsilon}} \mathbb{1}[\phi(x_i) \neq \hat{\phi}(x_i)]. \tag{6}$$

High values of these measures are desired, as that would imply that the $\mathbb{C}_{\epsilon}$ contains diverse classifiers which can identify more number of datapoints that have a contradictory decision for a given value of ϵ. We also report the time it takes to compute the set of classifiers $\mathbb{C}_{\epsilon}$.

Fairness benchmark. For results on fairness in the ambiguous regions, we compare our method given by Problem P5 using $\mathbb{C}_{\epsilon,\theta_{best}}$, picking classifiers uniformly at random from $\mathbb{C}_{\epsilon,\theta_{best}}$, the most accurate classifier and a traditional fair classifier. We chose one traditional fair method as a baseline, as Zafar et al. [29] show comparison to other approaches and get similar results. Its formulation ([29] and [1]) is given as follows,

$$\begin{aligned} \text{minimize} \quad & -\frac{1}{|\mathcal{D}|}\sum_{(x,y)\in\mathcal{D}} \log p(y|x,\theta) + \lambda||\theta|| && \text{(P6)} \\ \text{subject to} \quad & \frac{1}{|\mathcal{D}_*|}\left|\sum_{(x,z)\in\mathcal{D}_*}(z-\bar{z})d_{\theta}(x_i)\right| < c, \end{aligned}$$

where $\mathcal{D}_*$ was set to datapoints with ground truth negative labels and ground truth positive labels for equalizing false positive rates (FPR) and false negative rates (FNR), respectively. z represents the value of the sensitive feature and c represents the allowed correlation between z and the decision boundary, d_{θ}.

We train accurate classifiers by solving minimize $\;-\frac{1}{|\mathcal{D}|}\sum_{(x,y)\in\mathcal{D}}\log p(y|x,\theta) + \lambda||\theta||$ for different λ. Logistic regression loss was used to train all the classifier. More details such as ranges for the hyper parameter search values, seeds, specifications of the machines used and other training details are included in the appendix.

Metrics for fairness. We assume a binary valued sensitive attribute and report a signed difference of FPR and FNR between the unprotected and the protected group for the sensitive feature z.

$$\text{unfairness-FPR} = FPR_{z=1} - FPR_{z=0}, \tag{7}$$

$$\text{unfairness-FNR} = FNR_{z=1} - FNR_{z=0} \tag{8}$$

We present these numbers for the overall data, for the unambiguous regions, i.e., where all the classifiers give unanimous decisions, and for the ambiguous regions. We also report the accuracies. We aim to achieve low disparity in group error rates in the ambiguous regions, while achieving an accuracy similar to the most accurate classifier.

4.4 Synthetic Experiments

In this section, we answer the evaluations questions using the synthetic dataset.

Q1: Ambiguous regions coverage and speed. We compared our methods, Dsc-Approx and Amb-Approx, of identifying the ambiguous regions with Dsc-Exact and Amb-Exact. Table 2 reports the time it took to compute the ambiguous regions as well as the metrics described in Section 4.2. The results demonstrates that our methods are comparable or even better in coverage of the ambiguous regions on the test data, while being up to four orders of magnitude faster.

Q2: Accuracy fairness trade-off. We compare our method with the benchmarks described in Section 4.2. The results in Table 1 demonstrate that:

Existing fairness methods sometime achieves overall fairness at the expense of a significant decrease in accuracy. Additionally, overall fairness is achieved by being biased towards different groups for different types of errors, i.e., ones in the unambiguous vs ambiguous regions. On the other hand, our method is effective in removing unfairness in the ambiguous regions and ignoring the unfairness in the unambiguous regions, as desired. Our method also achieves accuracy similar to the most accurate classifiers.

Q3: Robustness to noisy data. In order to demonstrate the sensitivity of existing fairness methods towards noise, we flipped the ground truth labels of 0.0% to 20% of the datapoints uniformly at random. Figures 2 and 3 present our findings. We compare an accurate classifier, a fair classifier and our method equalizing FPR using Amb-Approx. The key takeaways are as follows: In an effort to equalize all errors, existing fairness methods are affected by label noise and end up classifying a significant number of datapoints in the wrong class, as hypothesized in the introduction. In contrast, our method is robust to noise as it identifies similar regions as ambiguous for varying level of noise. Secondly, this experiment also confirms our hypothesis, by showing that the ambiguous region coincide with regions with predominantly high model uncertainty, i.e., the sparse clusters.

4.5 Evaluation on Real-World Datasets

In this section, we answer our evaluation questions using two real-world datasets.

Q1: Ambiguous regions coverage and speed. We identify the datapoints with ambiguous decisions using Dsc-Approx, given by Problem P3 and Amb-Approx, given by Problem P4, for the same value of ϵ. We also tried Dsc-Exact and Amb-Exact, however after several hours of computations they still did not yield any results. So, we compare the results of our two proposals, using $\hat{\alpha}$ metric given by Equation 6. Takeaways remain similar for $\hat{\delta}$, given by Equation 5.

For the Compas data, our method Dsc-Approx and Amb-Approx categorized 0.12 and 0.5 of the datapoints as having an ambiguous decision, respectively. While for the SQF dataset, 0.12 and 0.53 of the datapoints were identified as having an ambiguous decision by Dsc-Approx and Amb-Approx, respectively. It is noteworthy that Amb-Approx identifies more datapoints as ambiguous. This is due to the fact that with Amb-Approx we train one classifier per training datapoint, i.e., we perform a more exhaustive search for the classifiers that exhibit predictive multiplicity. This process, however, takes a longer time. Hence, there is a trade-off between the speed and effectiveness for both the proposed methods of identifying the ambiguous regions.

Q2: Accuracy fairness trade-off. Similar to the synthetic dataset, we compare our method of equalizing group error rates (FPR and FNR) in the ambiguous regions, identified by Dsc-Approx and Amb-Approx, with three benchmarks described in Section 4.2. The takeaways from results presented in Tables 3 and 4 are the following. Existing fair classifiers that focus on equalizing overall error have high unfairness in the ambiguous regions in most cases, which confirms our hypothesis. Although these classifiers achieve fairness in the overall data, they sometimes result in a significant drop in accuracy. Additionally, in many cases, existing fair classifiers achieve overall fairness by being unfair to different groups in the ambiguous vs unambiguous regions.

In comparison, our method that only equalizes errors in the ambiguous regions, in most cases, provides the fairest solution in the ambiguous regions while achieving a comparable accuracy to the most accurate classifier.

In a few cases where our approach is not the only best solution, it provides additional benefits, e.g., in one case our solution is equally fair in the ambiguous region compared to the accurate classifier (cf. Table 4). However, our method assigns decisions to datapoints in the ambiguous regions stochastically. So, in practice, most datapoint in the ambiguous region have a non-zero probability to be in the favorable class. This is desirable over a deterministic decision, since there is ambiguity in decisions for these datapoints. In another case, Table 3, selecting classifiers uniformly at random is 1.6% more fair on the *test data*. However, our solution is still 90% and 18% better than the benchmark fair classifier and the accurate classifier, which are the current standards.

Table 4: [SQF] Signed differences in FPR/FNR

	Unfairness			Accuracy
	total	unamb	amb	
Acc.	-0.28/0.12	-0.29/0.13	-0.07/**0.017**	0.75
Fair	0.04/0.02	0.02/0.03	0.07/-0.15	0.65/0.71
Uni-P3	-0.28/0.12	-0.29/0.13	-0.05/**0.014**	0.75 / 0.75
Our-P3	-0.28/0.11	-0.29/0.13	**-0.02**/**-0.017**	0.75/ 0.75
With P4				
Acc.	-0.28/0.12	-0.24/0.17	-0.25/**0.07**	0.75
Fair	0.04/0.02	-0.06/0.12	**0.15**/-0.08	0.65/0.71
Uni-P4	-0.27/0.14	-0.24/0.17	-0.25/0.09	0.74/ 0.74
Our-P4	-0.24/0.13	-0.24/0.17	-0.18/**0.07**	0.73/ 0.74

This table demonstrates effectiveness of our methods. Please refer to Section 4.5

5 RELATED WORK

Fairness in ML. In recent years a number of fairness methods and notions have been proposed for classification tasks [1, 4, 9, 10, 12, 13, 17, 18, 24, 26–30]. A family of these methods aim to enforce fairness across socially salient groups in the society that equalize 'total' errors e.g., [1, 13, 27, 28]. In contrast, we argue to focus only on the errors arising due to *model uncertainty*. We do so by building on existing work in predictive multiplicity.

Modeling uncertainty. Prior works on categorizing uncertainties have proposed to distinguish between aleatoric (irreducible) uncertainty and model (reducible) uncertainty[6, 14, 15]. A lot of works in machine learning have addressed this distinction in different subfields. Depeweg et al. [5] propose to decompose the two types of uncertainties using bayesian neural networks and latent variables. Kendall and Gal [16] consider this distinction in computer vision problems. McAllister [22] distinguish between the types of uncertainties in reinforcement learning problems.

We believe that we are the first ones to propose to distinguish between different types of uncertainties for fairness in predictive tasks.

Predictive multiplicity. In their seminal work, Breiman et al. [3] introduced the concept of the *Rashomon effect* in the context of model explanations. The Rashomon effect refers to the scenario where data admits multiple different models that yield similar accuracy. Breiman et al. [3] argue that one should not use the explanations of a single model to draw conclusions about the data and the prediction task at hand. Rashomon sets, defined as ϵ-set of models, i.e. those whose empirical training loss is within ϵ-loss of a baseline classifier, are used by Dong and Rudin [8], Fisher et al. [11] to study the problem of variable importance.

The notion of predictive multiplicity in a classification setting was introduced by Marx et al. [21]. They proposed mixed integer programming methods using non-convex loss functions to train classifiers which would yield predictive multiplicity for linear classifiers. We build on this work, and extend it by proposing tractable convex problem formulations which yield fast solutions, and work for both linear and non-linear classifiers.

There is a growing interest in predictive multiplicity due to its societal implications on algorithmic decision-making system. Bhatt et al. [2] look at it from a fairness perspective, and aim to find counterfactual accuracy of a classifier which would give a selected test datapoint favorable outcome. Specifically, they aim to find the minimum decrease in accuarcy, $\boldsymbol{\epsilon}$, that would give an individual a favorable outcome. Pawelczyk et al. [23] provide an upper bound for the costs of finding counterfactual explanations under predictive multiplicity. However, none of these works have made the connections between predictive multiplicity and model uncertainty.

6 CONCLUDING DISCUSSION

In this work, we propose that while designing fairness approaches one must account for the uncertainties of the prediction task at hand. Specifically, we argue that only the *errors* arising due to lack of knowledge about the best model or due to lack of data, i.e., the *epistemic errors* should be taken into account while designing fairness methods and errors due to inherent noise should be ignored. Our proposal stands in contrast to the current group fairness approach that aims to equalize 'total' errors. With this goal in mind, we build upon predictive multiplicity techniques to identify the regions with model uncertainty.

In addition, we propose convex and scalable formulations to find classifiers that exhibit predictive multiplicity, which are approximately equally effective compared to their non-convex counterparts, while being up to four orders of magnitude faster. We also propose convex formulations to equalize errors arising due to model uncertainty. Using synthetic and real-world datasets, we demonstrate that our methods are effective and more robust to label noise compared to existing group fairness methods.

Our key insight is that not all types of errors are equal and when improving parity in the error rates one must account for the type of uncertainty inducing the error. We believe that this insight and our predictive multiplicity methods open new avenues for research on how to account for uncertainties when designing fair machine learning methods.

ACKNOWLEDGEMENTS

This research was supported in part by a European Research Council (ERC) Advanced Grant for the project “Foundations for Fair Social Computing”, funded under the European Union’s Horizon 2020 Framework Programme (grant agreement no. 789373). This work was also partly supported by the ERC Synergy Grant 610150 (imPACT).

REFERENCES

[1] Junaid Ali, Muhammad Bilal Zafar, Adish Singla, and Krishna P Gummadi. 2019. Loss-aversively fair classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 211–218.

[2] Umang Bhatt, Muhammad Bilal Zafar, Krishna Gummadi, and Adrian Weller. 2020. Counterfactual Accuracies for Alternative Models. *ICLR Workshop on Machine Learning in Real Life Workshop* (2020).

[3] Leo Breiman et al. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* 16, 3 (2001), 199–231.

[4] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *KDD*.

[5] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. 2018. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*. PMLR, 1184–1193.

[6] Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? Does it matter? *Structural safety* 31, 2 (2009), 105–112.

[7] Steven Diamond and Stephen Boyd. 2016. CVXPY: A Python-Embedded Modeling Language for Convex Optimization. *Journal of Machine Learning Research* 17, 83 (2016), 1–5.

[8] Jiayun Dong and Cynthia Rudin. 2019. Variable Importance Clouds: A Way to Explore Variable Importance for the Set of Good Models. *arXiv preprint arXiv:1901.03209* (2019).

[9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, and Omer Reingold. 2012. Fairness Through Awareness. In *ITCSC*.

[10] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *KDD*.

[11] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2019. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research* 20, 177 (2019), 1–81.

[12] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[13] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NIPS*.

[14] Stephen C Hora. 1996. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety* 54, 2-3 (1996), 217–223.

[15] Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning* 110, 3 (2021), 457–506.

[16] Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977* (2017).

[17] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. 2020. Fairness without Demographics through Adversarially Reweighted Learning. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 728–740.

[18] Preethi Lahoti, Krishna Gummadi, and Gerhard Weikum. 2019. Operationalizing Individual Fairness with Pairwise Fair Representations. *Proceedings of the VLDB Endowment* 13, 4 (2019), 506–518.

[19] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. Data and analysis for 'How we analyzed the COMPAS recidivism algorithm'. https://github.com/propublica/compas-analysis.

[20] Andrey Malinin. 2019. *Uncertainty Estimation in Deep Learning with Application to Spoken Language Assessment.* Ph.D. Dissertation. University of Cambridge.

[21] Charles T Marx, Flavio du Pin Calmon, and Berk Ustun. 2019. Predictive multiplicity in classification. *arXiv preprint arXiv:1909.06677* (2019).

[22] Rowan McAllister. 2017. *Bayesian learning for data-efficient control.* Ph.D. Dissertation. Department of Engineering, University of Cambridge.

[23] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 2020. On Counterfactual Explanations under Predictive Multiplicity. In *Conference on Uncertainty in Artificial Intelligence*. PMLR, 809–818.

[24] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware Data Mining. In *KDD*.

[25] Xinyue Shen, Steven Diamond, Yuantao Gu, and Stephen Boyd. 2016. Disciplined Convex-Concave Programming. *arXiv:1604.02639* (2016).

[26] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2239–2248.

[27] Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna P. Gummadi, and Adrian Weller. 2017. From Parity to Preference-based Notions of Fairness in Classification. In *NIPS*.

[28] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *WWW*.

[29] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*.

[30] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. 2013. Learning Fair Representations. In *ICML*.

A TRAINING DETAILS

In this section we explain the training details for our methods.

In order to train Dsc-Approx and Amb-Approx, presented in Section 4.1 of the paper, we used CPLEX library [25]. Dsc-Approx is give as follows,

$$\underbrace{\min_{\boldsymbol{\theta}} -\frac{1}{N}\sum_{\boldsymbol{x}_i, y_i} p(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})}_{\text{maximize accuracy}} \tag{P3}$$

$$\text{subject to: } \underbrace{\frac{1}{N}\sum_{\boldsymbol{x}_i}\max(0, d_{\boldsymbol{\theta}(\boldsymbol{x}_i)}d_{\boldsymbol{\theta}_{best}(\boldsymbol{x}_i)}) \leq \gamma}_{\text{limit agreement to } \boldsymbol{\theta}_{best}}$$

For synthetic dataset described in the paper we trained 1000 classifiers with $\gamma \in (1e-15, 2.0)$ picked linearly. For SQF dataset we also trained 1000 classifiers with $\gamma \in (0.0, 2.0)$ and for compas dataset we trained 1000 classifiers with $\gamma \in (0.0, 10.0)$ picked linearly.

In order to train the baselines mentioned in the experiment section of the paper, we trained 100 classifiers using logistic regression with L2 regularizer, minimize $-\frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y|\boldsymbol{x},\boldsymbol{\theta}) + \lambda||\boldsymbol{\theta}||$, with $\lambda \in (1e-1, 1)$, where $p(y=1|\boldsymbol{x},\boldsymbol{\theta}) = \frac{1}{1+\exp(-\boldsymbol{\theta}^T\boldsymbol{x})}$. We picked the λ that yielded the best accuracy on the validation set.

For traditional fairness methods given by,

$$\text{minimize} \quad -\frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y|\boldsymbol{x},\boldsymbol{\theta}) + \lambda||\boldsymbol{\theta}|| \tag{P6}$$

$$\text{subject to} \quad \frac{1}{|\mathcal{D}_*|}\left|\sum_{(\boldsymbol{x},z)\in\mathcal{D}_*}(z-\bar{z})d_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right| < c,$$

where $p(y=1|\boldsymbol{x},\boldsymbol{\theta}) = \frac{1}{1+\exp(-\boldsymbol{\theta}^T\boldsymbol{x})}$ and z is the sensitive attribute, same λ was used which we picked by training the accurate classifier. We trained 100 fair classifiers for each dataset by varying c values, which could be written as the product of correlation between different the sensitive attribute and $\boldsymbol{\theta}_{best}$ and multiplicative factor varying between zero and 1 [29], i.e., $c = t \cdot cov(\boldsymbol{\theta}_{best}, z)$. For synthetic dataset we used we use $t \in (0, 0.2)$ and for real world datasets $t \in (0.0, 1e-5)$. We train a pool of benchmark fair classifiers for varying values of c and a pool of accurate classifiers on 5 different shuffles of the data and then pick the fairest classifier and most accurate classifiers, respectively, for each shuffle from this pool.

We aggregated the results using these 5 seed values, [1122334455, 2211334455, 1133224455, 3322441155, 1122443355]. We used Intel(R) Xeon(R) CPU E7-8857 v2 @ 3.00GH with 48 cores to run all the experiments.

B PREDICTIVE MULTIPLICITY COMPARISON

In this section we show the visualization of the ambiguous regions with different methods introduced in the paper. Figure 5 shows ambiguous regions identified by the exact methods proposed by Marx et al. [21], Dsc-Exact and Amb-Exact, and our methods Dsc-Approx and Amb-Approx. The figure demonstrates that visually our methods identify similar regions with ambiguous results. In general, we also see that ambiguous regions are the more sparse regions of feature space, where decisions are difficult to make.

B.1 Results using nonlinear Classifiers

In this section we show the results using kernalized logistic regression to identify ambiguous regions, with Dsc-Approx. Figure 4 demonstrate the results.

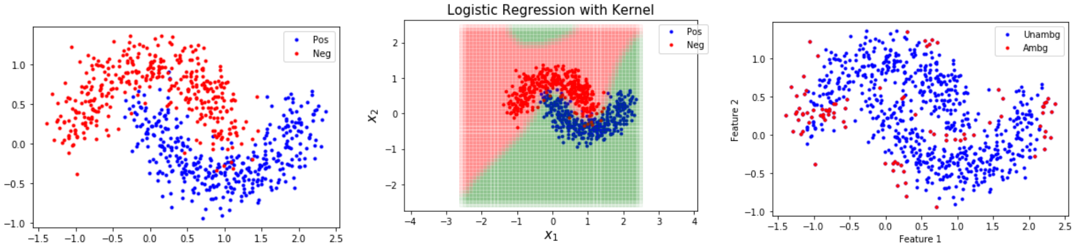

Figure 4: [Synthetic dataset-non-linear] The figure on the left shows the 2 moons dataset, the middle figure shows the best non-linear boundary with green regions classified as positive and red regions as negative and the one on the right shows the ambiguous regions identified using our method. The figure demonstrate that unlike Marx et al. [21] our methods can also be used to identify predictive multiplicity for non-linear classifiers.

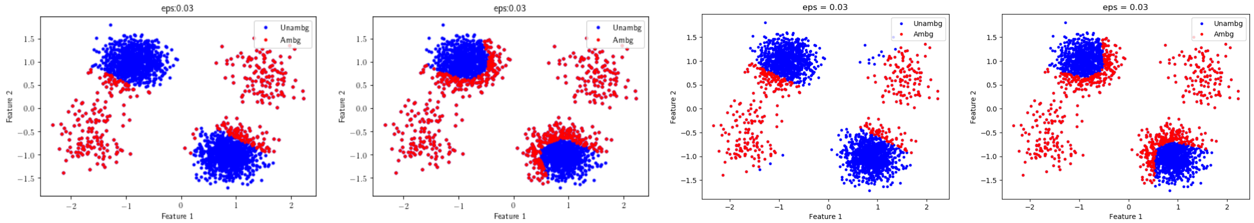

Figure 5: [Synthetic dataset] This figure shows the ambiguous regions identified by the four methods discussed in the paper. From left to right figures corresponds to Dsc-Exact, Amb-Exact, Dsc-Approx, Amb-Approx. It demonstrates that our methods identify similar ambiguous regions compared to the exact methods proposed by Marx et al. [21].