Rawlsian Fair Adaptation of Deep Learning Classifiers

Kulin Shah
Microsoft Research
Bangalore, India
kulin.shah98@gmail.com

Pooja Gupta
Indian Institute of Science
Bangalore, India
poojagupta@iisc.ac.in

Amit Deshpande
Microsoft Research
Bangalore, India
amitdesh@microsoft.com

Chiranjib Bhattacharyya
Indian Institute of Science
Bangalore, India
chiru@iisc.ac.in

ABSTRACT

Group-fairness in classification aims for equality of a predictive utility across different sensitive sub-populations, e.g., race or gender. Equality or near-equality constraints in group-fairness often worsen not only the aggregate utility but also the utility for the least advantaged sub-population. In this paper, we apply the principles of Pareto-efficiency and least-difference to the utility being accuracy, as an illustrative example, and arrive at the *Rawls classifier* that minimizes the error rate on the worst-off sensitive sub-population. Our mathematical characterization shows that the *Rawls classifier* uniformly applies a threshold to an ideal *score* of features, in the spirit of fair equality of opportunity. In practice, such a score or a feature representation is often computed by a black-box model that has been useful but unfair. Our second contribution is practical *Rawlsian fair adaptation* of any given black-box deep learning model, without changing the score or feature representation it computes. Given any score function or feature representation and only its second-order statistics on the sensitive sub-populations, we seek a threshold classifier on the given score or a linear threshold classifier on the given feature representation that achieves the *Rawls error rate* restricted to this hypothesis class. Our technical contribution is to formulate the above problems using ambiguous chance constraints, and to provide efficient algorithms for Rawlsian fair adaptation, along with provable upper bounds on the Rawls error rate. Our empirical results show significant improvement over state-of-the-art group-fair algorithms, even without retraining for fairness.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**; • **Theory of computation** → *Machine learning theory*; • **Social and professional topics** → Race and ethnicity; Gender.

KEYWORDS

Rawlsian Fairness; Fairness for Deep Learning Classifiers; Fair Adaptation

ACM Reference Format:
Kulin Shah, Pooja Gupta, Amit Deshpande, and Chiranjib Bhattacharyya. 2021. Rawlsian Fair Adaptation of Deep Learning Classifiers. 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, 14 pages. https://doi.org/10.1145/3461702.3462592

1 INTRODUCTION

Algorithmic decisions and risk assessment tools in real-world applications, e.g., recruitment, loan qualification, recidivism, have several examples of effective and scalable black-box models, many of which have received strong criticism for exhibiting racial or gender bias [6, 7]. Anti-discrimination laws and official public policy statements [1, 2, 27, 31] often explicitly demand that such black-box models be fairness-compliant.

Fair classification is an important problem in fairness-aware learning. Previous work on fair classification has largely focused on group-fairness or *classification parity*, which means equal or near-equal predictive performance on different sensitive sub-populations, e.g., similar accuracy [11, 15, 32], statistical parity [17], equalized odds [22], similar false positive rates [22]. A popular objective in group-fair classification is accuracy maximization subject to equal or near-equal predictive performance on different sensitive sub-populations.

What makes any fairness objective well founded? In this paper, we consider the basic principles of *Pareto-efficiency* and *least-difference*, following the work of Rawls on distributive justice [34–36]. A *Pareto-efficient* solution maximizes some aggregate utility over the entire population. Thus, it is not possible to improve the performance on any one sub-population without sacrificing performance on another sub-population beyond what a Pareto-efficient solution gives. The *least-difference* principle allows certain inequalities as fair, if reducing these inequalities either adversely affects the others or does not uplift the worst-off. For example, a group-fair solution that equalizes utilities across different sensitive sub-populations satisfies the least-difference principle vacuously because it has no inequality. Therefore, typical group-fair algorithms maximize accuracy subject to group-fairness constraints ensuring near-equal utilities for different sensitive sub-populations. However, none of the above solutions satisfy both the Pareto-efficiency and

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

ACM ISBN 978-1-4503-8473-5/21/05…$15.00
https://doi.org/10.1145/3461702.3462592

least-difference principles simultaneously. Do there exist solutions that satisfy both these principles simultaneously?

The most accurate classifier is the Bayes classifier that predicts using a uniform threshold of $1/2$ applied to the true outcome probability of every example. Previous work provides similar characterization of optimal group-fair classifiers as threshold classifiers on some ideal score [11–13, 32]. Do such threshold-on-score characterizations hold for Rawlsian fairness too?

Most black-box models in practice are often optimized for accuracy, and they output a risk score or a feature representation along with their prediction. Even when an existing model is found to be biased during a fairness audit, retraining a different fair model is difficult on proprietary or private training data. Algorithmic ideas to develop group-fair classifiers that maximize accuracy subject to group-fairness constraints require estimates of true outcome probability [11, 32], which are intractable with finite samples. Most post-processing methods for group-fairness also require estimates of the true outcome probabilities [22, 29, 33]. Secondly, existing group-fairness toolkit [11] can use a given risk score as an input feature and adapt for group-fairness but the resulting group-fair classifier does not predict using a threshold on that same risk score. Is fair adaptation possible without changing the risk scores or feature representations?

The cost or benefit of a decision for an individual depends on the protected attribute (e.g., race, gender, age) as well as the true class label. For example, an underprivileged person who qualifies for a loan and successfully repays, has a greater utility than a privileged person who qualifies for the same loan and repays, and also greater utility than another underprivileged person who qualifies but does not repay. Many group-fair classifiers with high and near-equal accuracy on different races or genders can have poor accuracy on either the positive or negative class therein. This typecasting is exacerbated when there is class imbalance within a group. Therefore, we consider the sensitive sub-populations defined by group memberships (e.g., race, gender, age) as well as true class labels.

1.1 Our results

In Section 3, we define the *Rawls classifier* as an optimal solution to an objective that simultaneously satisfies both the *Pareto-efficiency* and *least-difference* principles in fair classification. When the classification utility is quantified by accuracy, the *Rawls classifier* minimizes the error rate on the worst-off sensitive sub-population over all classifiers; we call its optimal value as the *Rawls error rate* (see Definition 3). The max-min objective is well-known in social choice theory for equitable distribution of goods [21, 39]; we provide a formulation so that it applies to fair classification. We mathematically characterize the *Rawls classifier* and the *Rawls error rate* (Theorem 6). Moreover, our additional observations in Subsection 3.2 show that the characterization of the *Rawls classifier* reveals interesting, non-trivial properties about its most disadvantaged sensitive sub-populations.

Our mathematical characterization of the *Rawls classifier* shows that it uniformly applies a threshold to a certain ideal *score* of features, whose description requires the underlying data distribution explicitly. We give a description of the *Rawls classifier* as a threshold on an ideal *score* function given by a convex combination of signed *unveil functions*, which quantify the likelihood of an individual belonging to a sensitive sub-population given only the unprotected attributes. Computing these unveil functions from a finite sample drawn randomly from an arbitrary underlying distribution is intractable, a familiar obstacle encountered for exact implementation of the accuracy maximizing Bayes classifier and the optimal group-fair classifier characterized in previous work [11, 32].

For Rawlsian *fair adaptation* of deep learning classifiers in practice, we consider the following formulation. Given a non-ideal score function or a feature map computed by a black-box model, and only its second-order statistics on the sensitive sub-populations, we seek a threshold classifier on the given non-ideal score or a linear threshold classifier on the given feature map that achieves the Rawls error rate in this restricted hypothesis class. In Section 4, we formulate the above problems using ambiguous chance constraints, and provide efficient algorithms for fair adaptation, along with provable upper bounds on the Rawls error rate in the restricted setting. In Subsection 4.2, we show that, when the feature map distributions conditioned on each sensitive sub-population are Gaussian and we seek a linear threshold classifier, then we can provably and efficiently achieve the restricted Rawls error rate (Theorem 13).

In Section 5, we show that our Rawlsian fair adaptation formulation above is readily applicable to any black-box model that computes a score function or a feature representation. For example, we train a model to maximize classification accuracy on a standard dataset used in text classification for toxicity. Our Rawlsian adaptation using its label-likelihood scores and feature representation does not require retraining a different fair model, and shows a significant improvement in the error rates on the worst-off sensitive sub-populations. We also show a similar improvement over real-world and synthetic data sets, when compared against best known post-processing fairness methods [29] and group-fair classifiers [11] as our baselines.

1.2 Related work

In the context of fair classification, recent work by Hashimoto et al. [23] studies Rawlsian fairness for empirical risk minimization, and observes that Rawlsian fairness prevents disparity amplification over time, which may be unavoidable if we insist on near-equal group-wise performance as a group-fairness constraint. Recent work has also looked at Rawlsian theory to study the veil of ignorance and inequality measurements [20, 24, 25, 38], contextual bandits [28], fair meta-learning [42], envy-free classification [26].

Recent work has also looked at Rawlsian theory to study fairness in different settings of machine learning such as the *veil of ignorance*, moral luck-vs-desert, difference principle etc. to study fair equality of opportunity, inequality measurements, bandit problems [20, 24, 25, 38]. Our max-min objective is a *prioritarian* objective based only on the Pareto-efficiency and the least-difference principles from distributive justice.

Previous work has also proposed different approaches for post-processing to achieve group-fairness on black-box models [3, 8, 18, 19, 22, 29]. However, they do not address fair adaptation of black-box deep learning models without changing their scores or feature representation similar to our work.

2 NOTATION

Let $\mathcal{X}$ be the space of input features, $\{0, 1\}$ be the binary class labels, and $[p] = \{1, 2, \ldots, p\}$ be the set of protected attributes, e.g., race, gender, age. Any input data distribution corresponds to a joint distribution $\mathcal{D}$ on $\mathcal{X} \times \{0, 1\} \times [p]$. Let (X, Y, Z) denote a random element of $\mathcal{X} \times \{0, 1\} \times [p]$ drawn from the joint distribution $\mathcal{D}$. Let p_{ij} denote $\Pr(Y = i, Z = j)$.

Definition 1. *For any $i \in \{0, 1\}, j \in [p]$, define the sensitive sub-population $S_{ij} \subseteq \mathcal{X} \times \{0, 1\} \times [p]$ as $S_{ij} = \{(x, i, j) : x \in \mathcal{X}\}$.*

The utility of a decision often depends on the protected attribute as well as the true label. This is implicitly considered in metrics such as false positive rates over groups. Our definition of sensitive sub-populations makes this more explicit.

Definition 2. *For any $i \in \{0, 1\}, j \in [p]$, define the unveil function $\eta_{ij} : \mathcal{X} \to \mathbb{R}_{\geq 0}$ of the sensitive sub-population S_{ij} as*

$$\eta_{ij}(x) = \Pr(Y = i, Z = j \mid X = x).$$

We define the normalized unveil function $u_{ij} : \mathcal{X} \to \mathbb{R}_{\geq 0}$ as

$$u_{ij}(x) = \eta_{ij}(x)/p_{ij}, \quad \text{where } p_{ij} = \Pr(Y = i, Z = j).$$

We call u_{ij} as the normalized unveil function because

$$\mathbb{E}_X\left[\frac{\eta_{ij}(X)}{p_{ij}}\right] = 1, \qquad \text{for all } i \in \{0, 1\}, j \in [p].$$

Let $\mathcal{D}_{ij}$ denote the conditional distribution for X given the class label $Y = i$ and the protected attribute $Z = j$, and X_{ij} denote a random element of $\mathcal{X}$ drawn from $\mathcal{D}_{ij}$.

As is common in fairness literature, we denote a binary classifier by a function $f : \mathcal{X} \to \{0, 1\}$. Note that this subsumes both group-aware as well as group-blind classifiers depending on whether the protected attributes also appear in $\mathcal{X}$ or not.

3 THE RAWLS CLASSIFIER

A natural way to measure the cost of a binary classifier f on the sensitive sub-population S_{ij} is by its error rate $r_{ij}(f) = \Pr(f(X) \neq Y \mid Y = i, Z = j)$.

Here are two well-known basic principles of efficiency and fairness from distributive justice and social choice theory, when we consider classification accuracy as utility.

(1) **Pareto-efficiency principle:** A classifier f is Pareto-efficient, if there exists some $\lambda \in \mathbb{R}^{2p}_{\geq 0}$ such that f minimizes $\sum_{ij} \lambda_{ij} r_{ij}(f)$ over all $f : \mathcal{X} \to \{0, 1\}$. As a consequence, it is not possible to improve the performance on any one sub-population without sacrificing performance on another sub-population.
(2) **Least-difference principle:** A classifier f satisfies the least-difference principle, if we have $|r_{ij}(g) - r_{kl}(g)| < |r_{ij}(f) - r_{kl}(f)|$, for any classifier g and two sensitive sub-populations S_{ij} and S_{kl}, then either $r_{ab}(g) > r_{ab}(f)$, for some a, b, or $\max_{(a,b)} r_{ab}(g) \geq \max_{(a,b)} r_{ab}(f)$. In other words, we cannot uplift the worst-off and reduce existing inequality without adversely affecting others.

Let $\hat{f}$ be the Bayes classifier that maximizes accuracy. Then $\hat{f}$ minimizes the total error $\sum_{ij} p_{ij} r_{ij}(f)$ among all $f : \mathcal{X} \to \{0, 1\}$, and hence it is Pareto-efficient. However, the Bayes classifier $\hat{f}$ can violate the least-difference principle.

A group-fair classifier that equalizes false positive rates $r_{0j}(f)$'s for all j, or false negative error rates $r_{1j}(f)$'s for all j, vacuously satisfies the least-difference principle, assuming we restrict ourselves to either only S_{0j}'s or only S_{1j}'s as sensitive sub-populations, respectively. However, a group-fair solution or even a solution that maximizes accuracy subject to near-equality group-fairness constraints is not necessarily Pareto-efficient. Insisting on equality or near-equality often violates the principle of Pareto-efficiency when the only way to achieve equality makes it worse but equal for individual sensitive sub-populations.

Now we state our objective that defines the Rawls classifier, and we will show later in this section that it satisfies both the Pareto-efficiency and least-difference principles.

Definition 3. *Given any joint distribution $\mathcal{D}$ on $\mathcal{X} \times \{0, 1\} \times [p]$, we define the Rawls classifier and the Rawls error rate, respectively, as*

$$f^* = \operatorname*{argmin}_{f : \mathcal{X} \to \{0,1\}} \max_{i \in \{0,1\}, j \in [p]} r_{ij}(f),$$
$$r^* = \min_{f : \mathcal{X} \to \{0,1\}} \max_{i \in \{0,1\}, j \in [p]} r_{ij}(f),$$

where $r_{ij}(f) = \Pr(f(X) \neq Y \mid Y = i, Z = j)$, which is the error rate of f on the sensitive sub-population S_{ij}.

It is easy to see that the Rawls classifier satisfies the least-difference principle. Any classifier g that reduces existing inequalities from f^* with adversely effects, i.e., keeping $r_{ij}(g) \leq r_{ij}(f)$, for all i, j, must either leave the maximum error rate untouched. Otherwise, it would contradict f^* being the minimizer of $\max_{(i,j)} r_{ij}(f)$.

On the other hand, Pareto-efficiency of the Rawls classifier requires a short proof (see Subsection 3.1). Interestingly, this also helps in characterizing it as a threshold classifier on an *ideal* score function.

Remark: Under reasonable assumptions on utility functions and choice spaces in social choice theory, it is known that any solution that satisfies Pareto-efficiency and least-difference principles simultaneously must actually be a solution that maximizes the minimum utility across participants [21, 39]. More specifically, it is a lex-min solution, i.e., if there exist multiple solutions that maximize the minimum utility across participants then a lex-min solution maximizes the second minimum utility among them, and then the third minimum utility and so on. However, in this paper, we focus only on maximizing the minimum utility, or equivalently, minimizing the maximum error rate.

We show that the error rate $r_{ij}(f)$ of any binary classifier f on its sensitive sub-population S_{ij} can be expressed as a weighted expectation of $f(X)$, weighted by the normalized unveil function $u_{ij}(X)$ of the sensitive sub-population S_{ij} (see Definition 2). As a consequence, each $r_{ij}(f)$ is a linear function of f, when the data distribution $\mathcal{D}$ is fixed. This is a known observation already used in previous work [11, 32].

Proposition 4. *For any binary classifier $f : \mathcal{X} \to \{0, 1\}$, its error rate $r_{ij}(f)$ on a sensitive sub-populations S_{ij} equals*

$$r_{ij}(f) = \begin{cases} \mathbb{E}_X\left[f(X)\, u_{0j}(X)\right], & \text{for } i = 0, \\ 1 - \mathbb{E}_X\left[f(X)\, u_{1j}(X)\right], & \text{for } i = 1, \end{cases}$$

where $p_{ij} = \Pr(Y = i, Z = j)$ and $u_{ij}(x)$ the normalized unveil function of S_{ij} (see Definition 2).

For any binary classifier f, we express its maximum error rate $r_{ij}(f)$ over all sensitive sub-populations S_{ij} as the maximum over all possible convex combinations of $r_{ij}(f)$'s. As a consequence, $\max_{ij} r_{ij}(f)$ is also a linear function of f, when the coefficients in the optimal convex combination and the underlying data distribution $\mathcal{D}$ are fixed.

Proposition 5. *For any binary classifier $f : \mathcal{X} \to \{0,1\}$, the maximum error rate over all sensitive sub-populations S_{ij} equals*

$$\max_{i\in\{0,1\}, j\in[p]} r_{ij}(f)$$
$$= \max_{\substack{\sum_{ij} c_{ij} \leq 1 \\ c_{ij}\geq 0,\ \forall ij}} \mathbb{E}_X \left[f(X) \left(\sum_{i\in\{0,1\}, j\in[p]} (-1)^i c_{ij} u_{ij}(X) \right) \right] + \sum_{j\in[p]} c_{1j}.$$

Proof of Proposition 4 and Proposition 5 are given in Appendix A.

3.1 Characterization of the Rawls classifier

Now we are ready to characterize the Rawls classifier. We characterize the Rawls classifier as a threshold classifier on an ideal score function that is expressed as a certain convex combination of signed, normalized unveil functions $u_{ij}(x)$'s. Moreover, we show that the non-zero coefficients in this convex combination actually indicate the maximally disadvantaged or vulnerable sensitive sub-populations S_{ij}'s.

Theorem 6. *Given any data distribution $\mathcal{D}$ on $\mathcal{X} \times \{0,1\} \times [p]$, there exist non-negative coefficients c^*_{ij}, for $i \in \{0,1\}$ and $j \in [p]$, satisfying $\sum_{i\in\{0,1\}, j\in[p]} c^*_{ij} = 1$, such that the Rawls classifier achieving the Rawls error rate is given by*

$$f^*(x) = \mathbb{1}\left(\sum_{j\in[p]} c^*_{1j} u_{1j}(x) - \sum_{j\in[p]} c^*_{0j} u_{0j}(x) \geq 0 \right).$$

and the Rawls error rate is equal to $r^* = \sum_{i\in\{0,1\}, j\in[p]} c^*_{ij} r_{ij}(f^*)$.

As an immediate corollary, we get Pareto-efficiency of the Rawls classifier because

$$f^* = \operatorname*{argmin}_{f:\mathcal{X}\to\{0,1\}} \sum_{i\in\{0,1\}, j\in[p]} c^*_{ij} r_{ij}(f).$$

Proof of Theorem 6 is given in Appendix A. Our proof is inspired by [32], who introduced similar techniques in the context of group-fairness. The main difference from previous work that uses similar techniques [11, 32] is that the coefficients are for each $i \in \{0,1\}$ and $j \in [p]$, instead of only $j \in [p]$, and moreover, they have a special meaning as we will show that the non-zero coefficients c^*_{ij} indicate the most disadvantaged sensitive sub-populations.

In Theorem 6, the indices of non-zero coefficients c^*_{ij}'s actually correspond to the sensitive sub-populations that attain the Rawls error rate r^*, and are therefore, the maximally disadvantaged or vulnerable sensitive sub-populations.

3.2 Properties of the Rawls classifier

An interesting corollary of the above characterization theorem is that for any Rawls classifier, the maximally disadvantaged sensitive sub-population cannot be unique, unless the Rawls classifier is trivial.

Corollary 7. *For any Rawls classifier, the sensitive sub-population S_{ij} that attains the Rawls error rate $r_{ij}(f^*) = r^*$ cannot be unique, unless the Rawls classifier f^* is trivial (i.e., all-zeroes or all-ones).*

Another interesting corollary of Theorem 6 is that for any Rawls classifier, there exist at least two sensitive sub-populations, one from each class, that both attain the Rawls error rate.

Corollary 8. *For any Rawls classifier f^*, there exist two sensitive sub-populations S_{0j} and S_{1k}, one from each class, that both of them attain the Rawls error rate of $r_{0j}(f^*) = r_{1k}(f^*) = r^*$, unless the Rawls classifier is trivial (i.e., all-zeroes or all-ones).*
Remark: For the above sub-populations, j and k need not be equal.

It is interesting to study the Rawls classifier even in the case when there are no protected attributes, and the sensitive sub-populations are the positive and negative classes, respectively. This is not standard for group-fairness but nevertheless relevant from a broader fairness perspective, i.e., when $p = 1$, the Rawls classifier minimizes the maximum of the false positive rate and the false negative rate. As we shall see, in this special case, Theorem 6 implies that the Rawls classifier applies a threshold to $\Pr(Y = 1 \mid X = x|)$, similar to but not the same as the well-known Bayes classifier.

Corollary 9. *For $p = 1$, the Rawls classifier f^* is given by $\mathbb{1}(\eta(x) \geq t)$, where $\eta(x) = \Pr(Y = 1 \mid X = x)$, and the threshold t is equal to*

$$t = \frac{(c_{01}/p_{01})}{(c_{11}/p_{11}) + (c_{01}/p_{01})}.$$

Proof of Corollary 7, Corollary 8 and Corollary 9 are given in Appendix A.

4 THE RESTRICTED RAWLS CLASSIFIER AND FAIR ADAPTATION

Theorem 6 shows a characterization of the Rawls classifier as a threshold classifier on an *ideal* score function, which can be defined using the unveil functions $\eta_{ij}(x)$ of the sensitive sub-populations and the optimal coefficients c^*_{ij}'s. These require the knowledge of the underlying distribution $\mathcal{D}$ explicitly, in addition to the optimal coefficients c^*_{ij}'s that are not easy to compute. Moreover, in practice, computing the unveil functions $\eta_{ij}(x)$'s from a finite sample drawn from the distribution $\mathcal{D}$ is intractable. This leads to a natural question of whether there exists a more practical definition of the Rawls classifier and the Rawls error rate, and efficient algorithms to achieve these.

In practice, we often have a non-ideal score function or a feature map, and the classifier we compute needs to be efficient, and comes from a restricted hypothesis class, e.g., threshold classifier on a score or linear threshold classifier on a feature map. However, we do not know the underlying data distribution $\mathcal{D}$ explicitly but only have access to certain statistics of the given score function or feature map on each sensitive sub-population. For simplicity, we assume that we are given reliable estimates of the second-order statistics (i.e., means and second moments) of the score function or feature map on all sensitive sub-populations. This is a reasonable assumption because the second-order statistics can be estimated efficiently from a small sample of points from the underlying data distribution $\mathcal{D}$. Assuming no additional knowledge of $\mathcal{D}$ beyond the second-order statistics, leads to the following definitions of the *set of restricted score-distributions* and the *restricted Rawls classifier.*

Definition 10. *For any set of means $M = \{\mu_{ij}\}_{i\in\{0,1\},j\in[p]}$ with μ_{ij}'s in $\mathbb{R}^d$ for all ij, and any set of covariance matrices $V = \{\Sigma_{ij}\}_{i\in\{0,1\},j\in[p]}$ with $\Sigma_{ij} \in \mathbb{R}^{d\times d}$ for all ij, define the set of restricted score-distributions $\mathcal{R}_{MV}$ as all score-distribution pairs $(s, \mathcal{D})$ such that $s : \mathcal{X} \to \mathbb{R}^d$ is any score function (for $d = 1$) or any feature map (for $d \geq 2$), and $\mathcal{D}$ is any distribution on some $\mathcal{X} \times \{0,1\} \times [p]$ such that*

$$\mathbb{E}_{X_{ij}}\left[s(X_{ij})\right] = \mu_{ij}, \text{ and}$$
$$\mathbb{E}_{X_{ij}}\left[\left(s(X_{ij}) - \mu_{ij}\right)\left(s(X_{ij}) - \mu_{ij})^T\right)\right] = \Sigma_{ij},$$
$$\forall i \in \{0,1\}, j \in [p],$$

where (X, Y, Z) is a random sample from distribution $\mathcal{D}$, and X_{ij} is a random sample X conditioned on $Y = i, Z = j$.

Definition 11. *For any set of restricted score-distributions $\mathcal{R}_{MV}$ as in Definition 10, and any hypothesis class $\mathcal{F}$ of classifiers $f : \mathbb{R}^d \to \{0,1\}$, define the restricted Rawls classifier and the restricted Rawls error rate, respectively, as*

$$F^* = \operatorname*{argmin}_{f\in\mathcal{F}} \max_{i\in\{0,1\},j\in[p]} R_{ij}(f),$$
$$R^* = \min_{f\in\mathcal{F}} \max_{i\in\{0,1\},j\in[p]} R_{ij}(f),$$
$$\text{where } R_{ij}(f) = \sup_{(s,\mathcal{D})\in\mathcal{R}_{MV}} \Pr\left(f(s(X)) \neq Y \mid Y = i, Z = j\right)$$
$$= \sup_{(s,\mathcal{D})\in\mathcal{R}_{MV}} \Pr\left(f(s(X_{ij})) \neq Y\right).$$

Restricted Rawls setting described above restricts the classifier f to be from a given hypothesis class but it is also relaxation in the sense that the distribution $\mathcal{D}$ is allowed to vary as long as the second-order statistics of $s(X_{ij})$'s are fixed.

4.1 Fair Adaptation of Threshold (FAT) in the restricted Rawls setting

In Theorem 12, we characterize the restricted Rawls classifier and give a constructive, algorithmic proof for finding it by formulating the problem using ambiguous chance constraints. As a result, we are able to take any existing threshold classifier on some score function, collect the second-order statistics of its score function on all sensitive sub-populations, and efficiently adapt it to a restricted Rawls classifier that we call as Fair-Adapted Threshold (FAT) classifier. Note than many existing classifiers optimized for accuracy and group-fairness come with their own score functions, and we can efficiently adapt their thresholds.

Theorem 12. *For any set of restricted score-distributions $\mathcal{R}_{MV}$ for $d = 1$ given by means μ_{ij}'s and variances σ_{ij}^2's (as in Definition 10), and the hypothesis class $\mathcal{F}$ of threshold classifiers $f_b(x) = \mathbb{1}(s(x) \geq b)$, for the underlying score function, the corresponding restricted Rawls classifier is given by $F^*(x) = f_{b^*}(x) = \mathbb{1}(s(x) \geq b^*)$ where the threshold b^* is equal to*

$$b^* = \mu_{1j^*} - \sigma_{1j^*}\sqrt{\frac{\mu_{1j^*} - \mu_{0j^*}}{\sigma_{1j^*} + \sigma_{0j^*}}}$$
$$= \mu_{0j^*} + \sigma_{0j^*}\sqrt{\frac{\mu_{1j^*} - \mu_{0j^*}}{\sigma_{1j^*} + \sigma_{0j^*}}},$$
$$\text{where} \qquad j^* = \operatorname*{argmin}_{j\in[p]} \frac{\mu_{1j} - \mu_{0j}}{\sigma_{1j} + \sigma_{0j}}.$$

The algorithmic version of Theorem 12, that first estimates μ_{ij} and σ_{ij} using a finite sample from $\mathcal{D}$ and then computes b^* is what we call as Fair Adaptation of Threshold (FAT). We call the corresponding classifier f_{b^*} as the *Fair-Adapted Threshold (FAT) Classifier.*

Proof of Theorem 12 is given in Appendix B.

4.2 Fair Linear-Adaptation of Thresholds (FLAT) in the restricted Rawls setting

In this section, we consider the problem of finding fair adaptation in the restricted Rawls setting where we are a given score or feature map $s : \mathcal{X} \to \mathbb{R}^d$ as a black box, and we have its second order statistics of $s(X_{ij})$, where X_{ij} is a random sample from the sensitive sub-populations S_{ij}. We seek a linear threshold classifier on the feature map $s(x)$, so as to achieve the restricted Rawls error rate. Another simplifying assumption we make is to let the distributions $s(X_{ij})$'s be Gaussians with the given means μ_{ij}'s and covariance matrices Σ_{ij}'s. To compare with the remark made after Definition 11, the Gaussian assumption means that the distributions $s(X_{ij})$'s are completely characterized once their means and covariance matrices are known.

In absence of a given feature map, we can also use the basic features, if $\mathcal{X} \subseteq \mathbb{R}^d$. Also note that the even group-fair algorithms that implement Gaussian Naive Bayes (e.g., Meta Fair by Celis et al. [11]) also estimate second-order statistics of the data. The second-order statistics can be efficiently estimated from a small sample, in contrast with the unveil functions $\eta_{ij}(X)$ that are known to be intractable from a finite sample.

Theorem 13. *Let $\mathcal{R}_{MV}$ be any restricted set of score-distributions $\mathcal{R}_{MV}$ given by the means $\mu_{ij} \in \mathbb{R}^d$ and covariance matrices $\Sigma_{ij} \in \mathbb{R}^{d\times d}$ of $s(X_{ij})$, for a random sample X_{ij} from the sensitive sub-populations S_{ij}, with an underlying score map $s : \mathcal{X} \to \mathbb{R}^d$ and data distribution $\mathcal{D}$. Let the restricted hypothesis class $\mathcal{F}$ be linear threshold classifiers $f_{w,b}(x) = \mathbb{1}\left(w^T s(x) \geq b\right)$. Assuming that the distributions of $s(X_{ij})$ are Gaussians, the restricted Rawls classifier $f_{w,b}$ is given by solving the following optimization problem.*

$$w^* = \operatorname*{argmin}_{w} \max_{j} \left\|\Sigma_{1j}^{1/2} w\right\|_2 + \left\|\Sigma_{0j}^{1/2} w\right\|_2 \quad \text{subject to}$$
$$w^T(\mu_{1j} - \mu_{0j}) = 1, \ \forall j \in [p].$$

where the restricted Rawls error rate is given by

$$r^* = 1 - \Phi\left(\min_j \frac{(w^*)^T(\mu_{1j} - \mu_{0j})}{\left\|\Sigma_{1j}^{1/2} w^*\right\|_2 + \left\|\Sigma_{0j}^{1/2} w^*\right\|_2}\right),$$

where $\Phi(\cdot)$ *is the CDF of the standard normal variable* $N(0,1)$. *The optimal threshold for the Rawls classifier is given by*

$$\begin{aligned} b_* &= w_*^T \mu_{0j^*} + \Phi^{-1}(1-r^*) \left\|\Sigma_{0j^*}^{1/2} w_*\right\|_2 \\ &= (w_*)^T \mu_{1j^*} - \Phi^{-1}(1-r^*) \left\|\Sigma_{1j^*}^{1/2} w_*\right\|_2, \\ \textit{where} \quad j^* &= \underset{j}{\operatorname{argmin}} \frac{(w^*)^T(\mu_{1j} - \mu_{0j})}{\left\|\Sigma_{1j}^{1/2} w^*\right\|_2 + \left\|\Sigma_{0j}^{1/2} w^*\right\|_2} \end{aligned}$$

We call the restricted Rawls classifier obtained by algorithmic implementation of Theorem 13 as Fair Linear Adaptation of Thresholds (FLAT) classifier. In the experiments section, we consider two classifiers (1) FLAT-1, where we approximate $s(X_{ij})$'s by spherical Gaussians, and (2) FLAT-2, where we approximate $s(X_{ij})$'s by non-spherical Gaussians.

Proof of Theorem 13 is given in Appendix B.

5 EXPERIMENTS

5.1 Experimental setup

Our baselines for comparison include the given black-box model (neural network trained for maximizing accuracy), meta-fair classifier [11], Reject Option Classifier (ROC) [29]. We use maximum error rate across all sensitive sub-populations as the primary metric for evaluation. We use the acronym FAT for Fair Adaptation of Threshold. We use acronyms FLAT1 and FLAT2 to represent Fair Linear Adaptation of Threshold using spherical covariance matrix and complete covariance matrix, respectively.

In our experiments, we randomly split every dataset (except wikipedia talk page dataset) into training set (80 %) and testing set (20 %). We perform 10 repititions and report average statistics of all algorithms. We used a multilayer perceptron with 2 hidden layers to get the score for FAT algorithm or feature embedding for FLAT1 and FLAT2. We use 100 neurons in the first hidden layer and choose it from the range of 20 to 100 for the second hidden layer using accuracy. We used adam optimizer with batch size 128. We started the training with 0.005 learning rate with the step decay learning rate scheduler and optimized the parameters of the scheduler to maximize the accuracy.

5.2 Fair adaptation on real world datasets

In this section, we illustrate our Rawlsian fair adaptation of any given unfair classifier. We use three real world datasets to show experimental results of our proposed algorithms. The details of these datasets are given below.

- Wikipedia Talk Page Dataset [41]: In this dataset, the task is to predict whether a comment is toxic. Previous work has pointed out that training to maximize accuracy leads to unintended bias for comments containing terms 'gay', 'black' etc. ([16], [9]). The dataset has total 95,691 training examples and 31,867 test examples. We divide all the comments into six groups, that define our protected attributes. First five groups represent whether a comment contains the terms 'gay', 'islam', 'muslim', 'male', 'black', respectively, and the sixth group contains comments that contain none of the above. Note that a comment can be part of more than one subgroup.
- Adult Income Dataset [14]: In this dataset, the task is to predict whether an individual has income more than \$50K. The dataset has total 48,842 examples. Protected attributes in the dataset are gender and race.
- Bank Dataset [14]: In this dataset, the task is to predict whether a client will subscribe to a term deposit. The dataset

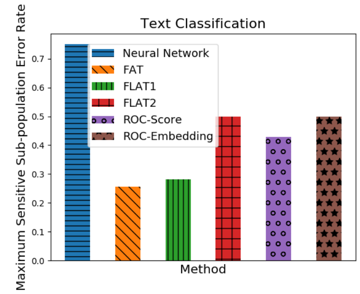

Figure 1: Comparison of maximum sub-population error rate on text classification

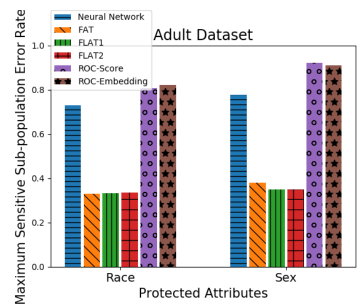

Figure 2: Adaptation of scores or feature embeddings on adult dataset

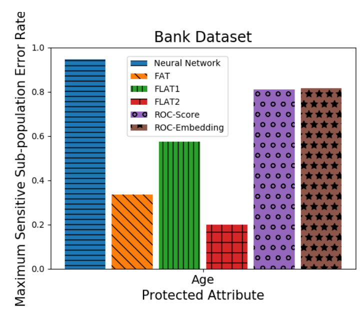

Figure 3: Adaptation of scores or feature embeddings on bank dataset

Sub-population	Mean	Variance	Number of Points
(0, 0)	(0, -2.5)	2	1900
(0, 1)	(5, 3)	1	100
(1, 0)	(0, 3)	2	1900
(1, 1)	(2, 5)	1	100

Table 1: Details of synthetic dataset 1. Sub-population (i, j) means population with label i and protected attribute j.

Sub-population	Mean	Variance	Number of Points
(0, 0)	(-5, 0)	2	1900
(0, 1)	(-1, -1)	1	100
(1, 0)	(5, 0)	2	1900
(1, 1)	(1, 1)	1	100

Table 2: Details of synthetic dataset 2. Sub-population (i, j) means population with label i and protected attribute j.

has total 45,211 examples. Protected attribute in the dataset is age.

We first show fair adaptation of non-linear feature embeddings and scores on the above mentioned datasets in Figure 1, Figure 2 and Figure 3. On each dataset, we first train a feedforward neural network to maximize overall accuracy. Then, we use second order statistics of scores (for FAT) or feature embeddings (for FLAT1, FLAT2) on each sensitive sub-population for Rawlsian fair adaptation. We use Reject Option Classifier (ROC) applied to the scores (ROC-Score) and feature embeddings (ROC-Embedding) of the neural network for comparison. Please note that ROC-Score and ROC-Embedding use the scores and feature embeddings for the entire data whereas FAT, FLAT1 and FLAT2 only use their second-order statistics on sensitive sub-populations.

In Figure 1, we see that FAT and FLAT1 give 10-15 % improvement in the maximum error rate over all sensitive sub-populations when compared to the baselines. However, FLAT2 has slightly higher maximum sensitive sub-population error rate than ROC-Score because our fair adaptation only uses second-order statistics on sensitive sub-populations. In Figure 2, we see that FAT, FLAT1 and FLAT2 achieve around 35-40 % improvement in the maximum error rate over all sensitive sub-populations when compared to ROC and the given neural network. On the bank dataset (figure 3), FAT and FLAT2 achieve around 40-50 % improvement in the maximum error rate over all sensitive sub-populations when compared to the baselines, whereas FLAT1 achieves around 20 % improvement over ROC and the given neural network.

5.3 Fair adaptation on synthetic datasets

We show the effectiveness of fair adaptation by showing quantitative and qualitative improvement on synthetic datasets. The synthetic datasets we consider have a binary-valued protected attribute (or two groups) and two classes. We use two-dimensional synthetic datasets for visualization of the decision boundaries given by different classifiers. For each fixed class label and protected attribute value, the features or examples are generated using a spherical Gaussian distribution. The parameters of these Gaussians and number of data points generated are given in Table 1 and Table 2. We compare the maximum error rate over all sensitive sub-populations for FLAT1 and Meta-fair classifier, and also the upper bound guarantee on this metric for FLAT1 computed by our algorithm in Figure 4. We see that FLAT1 outperforms state-of-the-art Meta-fair classifier and achieves around 70 % smaller value of the maximum error rate over sensitive sub-populations in synthetic dataset 1, and around 15 % smaller value in synthetic dataset 2. Figure 5, Figure 6, Figure 7 and Figure 8 show the different decision boundaries learnt by our

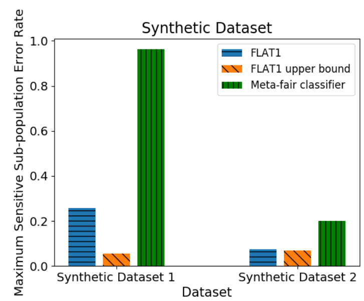

Figure 4: Comparison of maximum sensitive sub-population error rate for synthetic datasets

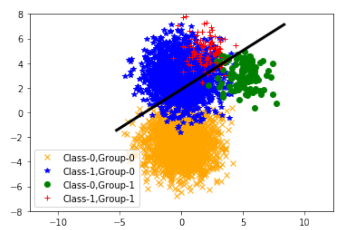

Figure 5: Decision boundary (black line) of FLAT1 on synthetic dataset 1

fair adaptation and Meta-fair. As Meta-fair classifier defines sub-population using only protected attribute and learns its decision boundary to equalize error rate on each sub-population defined only by the protected attribute, it almost ignores sub-population with label 0 and protected attribute value 1 in synthetic dataset 1. Moreover, Meta-fair classifier learns a non-linear complex classifier and still performs poorly. Because of Rawlsian fairness in our objective and defining the sub-population using class label and protected attribute, FLAT1 learns simple linear classifier and achieves lower maximum error rate across sensitive sub-populations. Even in synthetic dataset 2, FLAT1 learns a tilted decision boundary but Meta-fair learns a vertical decision boundary by ignoring sub-population with protected attribute value 1. Hence, we see that Rawlsian fairness objective and the definition of a sub-population using class label and protected attribute value helps us to do better than group-fair classifiers by only using second order statistics, and address typecasting of sub-populations caused by class imbalance.

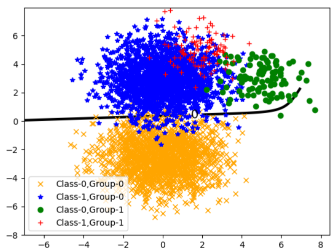

Figure 6: Decision boundary (black line) of Meta-fair classifier on synthetic dataset 1

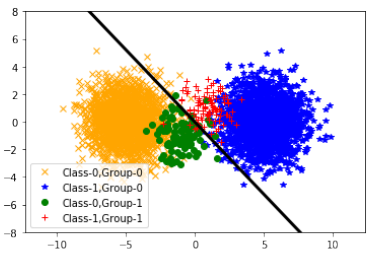

Figure 7: Decision boundary (black line) of FLAT1 on synthetic dataset 2

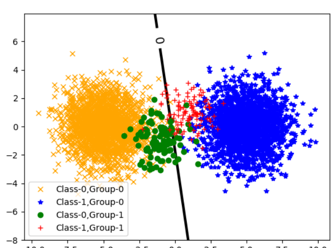

Figure 8: Decision boundary (black line) of Meta-fair classifier on synthetic dataset 2

5.4 Additional fair adaptation results

Here, we show additional fair adaptation results on two datasets. The details of these datasets are given below.

- COMPAS dataset [5]: In this dataset, the task is to predict recidivism from an individual's previous history (e.g. previous criminal history, prison time, etc.). The dataset has total 5,278 examples. Protected attributes in the dataset is gender and race.
- German dataset [14]: In this dataset, the task is to predict whether an individual is a good credit risk. The dataset has total 1,000 examples. Protected attributes in the dataset is gender and age.

In Figure 9 and Figure 10, we compare range of FPR and FNR of each group on Adult and Compas dataset. We use meta-fair

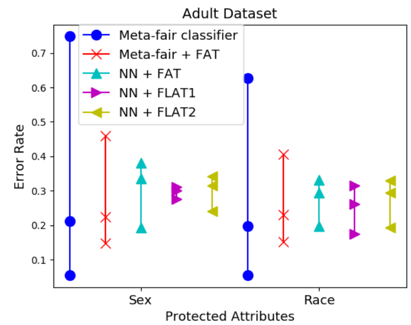

Figure 9: Comparison of range of FPR and FNR on Adult dataset. The lowest and highest point in a line denotes minimum and maximum value of FPR and FNR among all groups, respectively. The middle point among three points in a line denotes error rate.

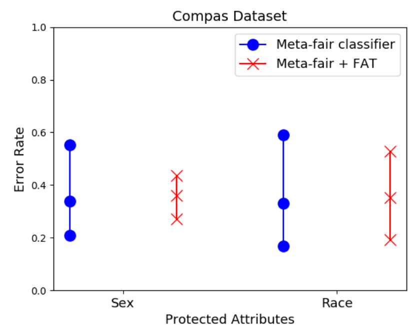

Figure 10: Comparison of range of FPR and FNR on Compas dataset. The lowest and highest point in a line denotes minimum and maximum value of FPR and FNR among all groups, respectively. The middle point among three points in a line denotes error rate.

classifier optimized to get near-equal FNR on each group as our baseline. As meta-fair classifier maximizes accuracy on combined data, it ignores one or more sensitive sub-population and ends up getting wide range of FPR and FNR values on each sensitive sub-population. However, in the proposed approaches, even though we are not directly optimizing FPR and FNR on each sensitive sub-population, we get smaller range compared to meta-fair classifier.

We show the results of Fair Adaptation of Threshold (FAT) algorithm in Figure 11 . We use second order statistics of each sub-populations of scores of meta-fair classifier and optimize over threshold to get group fair classifier using FAT algorithm. In Figure 11, we see that FAT algorithm improves maximum sensitive sub-population error rate by around 5-10 % in adult dataset and COMPAS dataset but in german dataset, FAT improves maximum sensitive sub-population error rate by around 40-50 % using only second-order statistics of each sub-populations.

We show experimental results of FLAT1 and FLAT2 on Adult dataset, COMPAS dataset and German Dataset in Figure 12. We use second order statistics of features of each sub-population for all

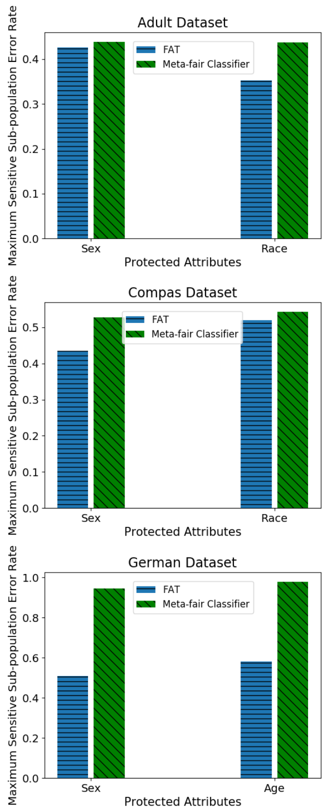

Figure 11: Comparison of FAT algorithm with meta-fair classifier on different datasets and protected attributes

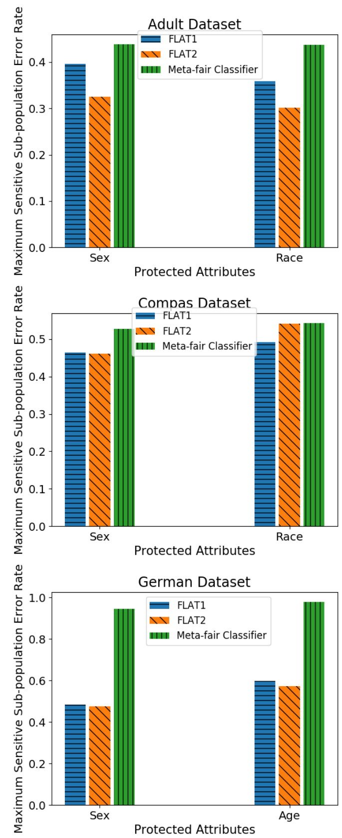

Figure 12: Comparison of FLAT1, FLAT2 algorithm with meta-fair classifier on different datasets and protected attributes

the datasets and learn linear classifier on the features using FLAT1 and FLAT2. For comparison, we learn meta-fair classifier using the entire data.

From figure 12, we see that FLAT1 and FLAT2 gets around 10-15 % lower maximum sensitive sub-population error rate on Adult and COMPAS dataset. In German dataset, FLAT1 and FLAT2 gets around 35-40 % lower maximum sensitive sub-population error rate. Thus, FLAT1 and FLAT2 achieves lower maximum sensitive sub-population error rate on all the datasets and all the protected attributes using only second order statistics of features of each sub-population.

Rawlsian fairness not only tries to minimize maximum sensitive sub-population error rate but in the process, it also decrease gap between error rate among all sensitive sub-populations. In figure 13, we show the comparison of error rate on each sub-population for neural network, FAT and FLAT2. For the figure, it is clear that the proposed algorithm achieves decreases the gap between error rates of sensitive sub-populations.

In figure 14, we show experimental results using False Positive Rate (FPR) and False Negative Rate (FNR) metric on Adult dataset. We compare our algorithm with Meta-fair classifier. We see that FLAT2 decreases gap in both metrics (FPR and FNR) between sensitive sub-populations.

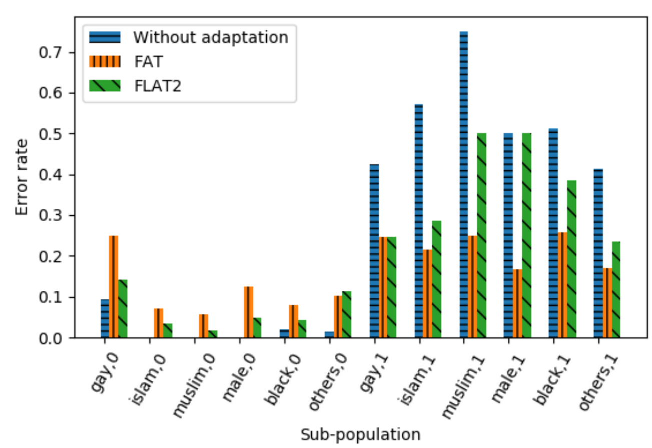

Figure 13: Group-wise class-wise error Rate on all sensitive sub-populations

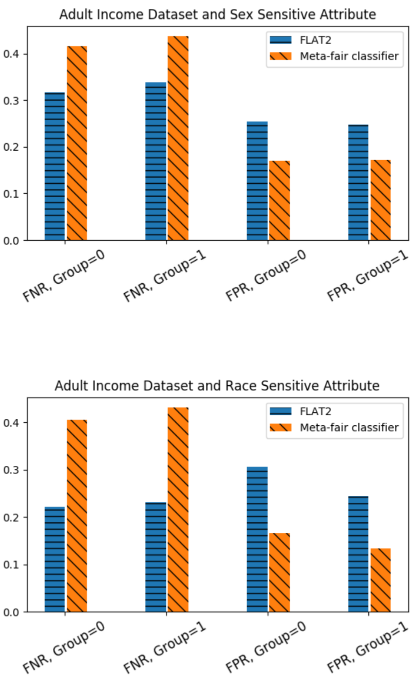

Figure 14: Comparison of FPR and FNR of FLAT2 algorithm with Meta-fair classifier on Adult dataset

6 CONCLUSION

Starting from the basic principles of Pareto-efficiency and least-difference in distributive justice, we characterize the Rawls classifier for minimizing the maximum error rate across all sensitive sub-populations. We also propose a principled approach for Rawlsian fair adaptation of black-box deep learning models that does not require retraining for fairness, while achieving significant improvement over state of the art group-fair baselines.

REFERENCES

[1] 1964. Civil rights act of 1964, Title VII.

[2] ACM. 2017. Statement on algorithmic transparency and accountability. https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

[3] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. 2018. A Reductions Approach to Fair Classification. In *ICML*.

[4] F. Alizadeh and D. Goldfarb. 2003. Second-order cone programming. *Mathematical Programming* 95, 1 (2003), 3–51.

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. COMPAS dataset. https://github.com/propublica/compas-analysis

[6] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: ThereâĂŹs software used across the country to predict future criminals, and itâĂŹs biased against blacks. *ProPublica* (May 2016).

[7] Solon Barocas and Andrew D Selbst. 2016. Big dataâĂŹs disparate impact. *California Law Review* (2016).

[8] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. *CoRR* abs/1707.00075 (2017). arXiv:1707.00075 http://arxiv.org/abs/1707.00075

[9] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*. 491–500.

[10] Stephen Boyd and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

[11] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. 2019. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In *FAT*.

[12] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *CoRR* abs/1808.00023 (2018).

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

[14] Dua Dheeru and Efi Karra Taniskidou. 2017. UCI machine learning repository. http://archive.ics.uci.edu/ml

[15] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. Compas risk scales: Demonstrating accuracy equity and predictive parity. (2016).

[16] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 67–73.

[17] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through awareness. In *ITCS*.

[18] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2018. Decoupled Classifiers for Group-Fair and Efficient Machine Learning. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*. PMLR, New York, NY, USA, 119–133.

[19] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD âĂŽ15)*. Association for Computing Machinery, 259âĂŞ268.

[20] Krishna P. Gummadi and Hoda Heidari. 2019. Economic Theories of Distributive Justice for Fair Machine Learning. In *Companion Proceedings of The 2019 World Wide Web Conference (WWW âĂŽ19)*. Association for Computing Machinery, 1301âĂŞ1302.

[21] Peter J Hammond. 1976. Equity, Arrow's conditions, and Rawls' difference principle. *Econometrica: Journal of the Econometric Society* (1976), 793–804.

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

[23] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness Without Demographics in Repeated Loss Minimization. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*. PMLR, 1929–1938.

[24] Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. 2018. Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making. In *Advances in Neural Information Processing Systems 31*. 1265–1276.

[25] Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. 2019. A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* âĂŽ19)*. Association for Computing Machinery, New York, NY, USA, 181âĂŞ190. https://doi.org/10.1145/3287560.3287584

[26] Safwan Hossain, Andjela Mladenovic, and Nisarg Shah. 2020. Designing Fairly Fair Classifiers Via Economic Fairness Notions. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 1559âĂŞ1569. https://doi.org/10.1145/3366423.3380228

[27] White House. 2016. Big data: A report on algorithmic systems, opportunity, and civil rights.

[28] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2016. Fair algorithms for infinite and contextual bandits. *arXiv preprint arXiv:1610.09559* (2016).

[29] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision Theory for Discrimination-Aware Classification. *2012 IEEE 12th International Conference on Data Mining* (2012), 924–929.

[30] Gert R. G. Lanckriet, Laurent El Ghaoui, Chiranjib Bhattacharyya, and Michael I. Jordan. 2002. A Robust Minimax Approach to Classification. *Journal of Machine Learning Research* 3 (2002), 555–582.

[31] Susan Magarey. 2004. The sex discrimination act 1984. *Australian Feminist Law Journal* (2004).

[32] Aditya Krishna Menon and Robert C. Williamson. 2018. The cost of fairness in binary classification. In *FAT*.

[33] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon M. Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. In *NIPS*.

[34] John Rawls. 1958. Justice as Fairness. *Philosophical Review* 67, 2 (1958), 164–194. https://doi.org/10.2307/2182612

[35] John Rawls. 1971. (1 ed.). Belknap.

[36] John Rawls. 2001. *Justice as Fairness: A Restatement*. Harvard University Press.

[37] S. Schaible. 1983. Fractional programming. *Zeitschrift für Operations-Research* 27, 1 (1983), 39–54.

[38] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD âĂŽ19)*. Association for Computing Machinery, New York, NY, USA, 2459âĂŞ2468.
[39] Steven Strasnick. 1976. Social choice and the derivation of Rawls's difference principle. *The Journal of Philosophy* 73, 4 (1976), 85–99.
[40] Philip Wolfe. 1976. Finding the Nearest Point in A Polytope. *Math. Program.* (1976), 128–149.
[41] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*. 1391–1399.
[42] Chen Zhao, Feng Chen, Zhuoyi Wang, and Latifur Khan. 2020. A Primal-Dual Subgradient Approach for Fair Meta Learning. *arXiv preprint arXiv:2009.12675* (2020).

A PROOF DETAILS FOR SECTION 3

A.1 Proof for Proposition 4

Proof. The error rate of f on the sensitive sub-population S_{ij} is given by $r_{ij}(f) = \Pr(f(X) \neq Y \mid Y = i, Z = j)$.

$$
\begin{aligned}
&\Pr(f(X) = 1 \mid Y = i, Z = j) \\
&= \int_{x \in \mathcal{X}} \mathbb{1}(f(x) = 1)\ \Pr(X = x \mid Y = i, Z = j)\, dx \\
&= \int_{x \in \mathcal{X}} f(x)\ \Pr(X = x \mid Y = i, Z = j)\, dx, \\
&\qquad\qquad\qquad \text{as } f : \mathcal{X} \rightarrow \{0,1\} \\
&= \int_{x \in \mathcal{X}} \frac{f(x)\ \Pr(Y = i, Z = j \mid X = x)\ \Pr(X = x)}{\Pr(Y = i, Z = j)} dx \\
&= \mathbb{E}_X\left[f(X)\ \frac{\eta_{ij}(X)}{p_{ij}}\right] \\
&= \mathbb{E}_X\left[f(X)\, u_{ij}(X)\right].
\end{aligned}
$$

where $\eta_{ij}(x) = \Pr(Y = i, Z = j \mid X = x)$ and $\eta_{ij}(x)$ and $u_{ij}(x)$ are the unveil function and the normalized unveil function, respectively, of S_{ij} as in Definition 2.

$$
\begin{aligned}
r_{0j}(f) &= \Pr(f(X) = 1 \mid Y = 0, Z = j) \\
&= \mathbb{E}_X\left[f(X)\frac{\eta_{0j}(X)}{p_{0j}}\right], \text{ and} \\
r_{1j}(f) &= \Pr(f(X) = 0 \mid Y = 1, Z = j) \\
&= 1 - \Pr(f(X) = 1 \mid Y = i, Z = j) \\
&= 1 - \mathbb{E}_X\left[f(X)\frac{\eta_{1j}(X)}{p_{1j}}\right] \\
&= 1 - \mathbb{E}_X\left[f(X)\, u_{1j}(X)\right].
\end{aligned}
$$

□

A.2 Proof for Proposition 5

Proof. For any binary classifier $f : \mathcal{X} \rightarrow \{0,1\}$,

$$
\begin{aligned}
&\max_{i \in \{0,1\}, j \in [p]} r_{ij}(f) \\
&= \max_{\substack{\sum_{ij} c_{ij} \leq 1 \\ c_{ij} \geq 0,\ \forall ij}} \sum_{i \in \{0,1\}, j \in [p]} c_{ij} r_{ij}(f) \\
&= \max_{\substack{\sum_{ij} c_{ij} \leq 1 \\ c_{ij} \geq 0,\ \forall ij}} \mathbb{E}_X\left[f(X)\left(\sum_{j \in [p]} c_{0j} u_{0j}(X) - c_{1j} u_{1j}(X)\right)\right] \\
&\qquad + \sum_{j \in [p]} c_{1j}, \qquad \text{using Proposition 4} \\
&= \max_{\substack{\sum_{ij} c_{ij} \leq 1 \\ c_{ij} \geq 0,\ \forall ij}} \mathbb{E}_X\left[f(X)\left(\sum_{i \in \{0,1\}, j \in [p]} (-1)^i c_{ij} u_{ij}(X)\right)\right] \\
&\qquad + \sum_{j \in [p]} c_{1j}.
\end{aligned}
$$

□

A.3 Proof for Theorem 6

Proof. We use Proposition 5 to express the Rawls error rate as minimizing a linear function over binary classifiers $f : \mathcal{X} \rightarrow \{0,1\}$, where the coefficients of the linear function come from solving an inner maximization. Then we relax our classifier to be real-valued in $[0,1]$, and use von Neumann's minimax theorem to swap the order of optimization. After the swap, it is easy to see that the optimal relaxed classifier is actually binary-valued, and it must be the Rawls classifier. By Definition 3 of the Rawls error rate,

$$
\begin{aligned}
r^* &= \min_{f : \mathcal{X} \rightarrow \{0,1\}} \max_{i \in \{0,1\}, j \in [p]} r_{ij}(f) \\
&= \max_{\substack{\sum_{ij} c_{ij} \leq 1 \\ c_{ij} \geq 0,\ \forall ij}} \sum_{i \in \{0,1\}, j \in [p]} c_{ij} r_{ij}(f) \\
&= \min_{f : \mathcal{X} \rightarrow \{0,1\}} \max_{\substack{\sum_{ij} c_{ij} \leq 1 \\ c_{ij} \geq 0,\ \forall ij}} \mathbb{E}_X\left[f(X)\left(\sum_{\substack{i \in \{0,1\} \\ j \in [p]}} (-1)^i c_{ij} u_{ij}(X)\right)\right] \\
&\quad + \sum_{j \in [p]} c_{1j}, \quad \text{using Proposition 5} \\
&\geq \min_{h : \mathcal{X} \rightarrow [0,1]} \max_{\substack{\sum_{ij} c_{ij} \leq 1 \\ c_{ij} \geq 0,\ \forall ij}} \mathbb{E}_X\left[h(X)\left(\sum_{\substack{i \in \{0,1\} \\ j \in [p]}} (-1)^i c_{ij} u_{ij}(X)\right)\right] \\
&\quad + \sum_{j \in [p]} c_{1j}, \ \text{using relaxation } h : \mathcal{X} \rightarrow [0,1]
\end{aligned}
$$

When $u_{ij}(x)$'s are known explicitly through the given distribution $\mathcal{D}$, by von Neumann's minimax theorem in variables $h(x)$ and c_{ij}'s, we get

$$r* \geq \max_{\substack{\sum_{ij} c_{ij} \leq 1 \\ c_{ij} \geq 0,\ \forall ij}} \min_{h:\mathcal{X}\to[0,1]} \mathbb{E}_X \left[h(X) \left(\sum_{\substack{i\in\{0,1\} \\ j\in[p]}} (-1)^i c_{ij} u_{ij}(X) \right) \right] + \sum_{j\in[p]} c_{1j},$$

$$= \max_{\substack{\sum_{ij} c_{ij} \leq 1 \\ c_{ij} \geq 0,\ \forall ij}} \mathbb{E}_X \left[h_c(X) \left(\sum_{\substack{i\in\{0,1\} \\ j\in[p]}} (-1)^i c_{ij} u_{ij}(X) \right) \right] + \sum_{j\in[p]} c_{1j},$$

where $h_c(x) = \mathbb{1}\left(\sum_{j\in[p]} (-1)^i c_{ij} u_{ij}(x) \leq 0\right)$.

$$\begin{aligned} r* &\geq \min_{f:\mathcal{X}\to\{0,1\}} \max_{\substack{\sum_{ij} c_{ij} \leq 1 \\ c_{ij} \geq 0,\ \forall ij}} \mathbb{E}_X \left[f(X) \left(\sum_{\substack{i\in\{0,1\} \\ j\in[p]}} (-1)^i c_{ij} u_{ij}(X) \right) \right] \\ &\quad + \sum_{j\in[p]} c_{1j}, \quad \text{because } h_c : \mathcal{X} \to \{0,1\} \\ &= \min_{f:\mathcal{X}\to\{0,1\}} \max_{i\in\{0,1\}, j\in[p]} r_{ij}(f), \quad \text{using Proposition 5} \\ &= r^*, \end{aligned}$$

Therefore, all the above inequalities must be equalities. Moreover, let $c^* = \left(c^*_{ij}\right)_{i\in\{0,1\}, j\in[p]}$ be the optimal solution of the above. Hence, the Rawls classifier is given by

$$\begin{aligned} f^*(x) = h_{c^*}(x) &= \mathbb{1}\left(\sum_{j\in[p]} (-1)^i c^*_{ij} u_{ij}(x) \leq 0 \right) \\ &= \mathbb{1}\left(\sum_{j\in[p]} c^*_{1j} u_{1j}(x) - \sum_{j\in[p]} c^*_{0j} u_{0j}(x) \geq 0 \right), \end{aligned}$$

and the Rawls error rate is equal to

$$\begin{aligned} r^* &= \sum_{i\in\{0,1\}, j\in[p]} c^*_{ij} r_{ij}(f^*) \\ &= \sum_{i\in\{0,1\}, j\in[p]} c^*_{ij} r_{ij}(h_{c^*}) \\ &= \mathbb{E}_X \left[h_{c^*}(X) \left(\sum_{i\in\{0,1\}, j\in[p]} (-1)^i c^*_{ij} u_{ij}(X) \right) \right] + \sum_{j\in[p]} c^*_{1j} \\ &= \mathbb{E}_X \left[\min\left\{ 0, \sum_{i\in\{0,1\}, j\in[p]} (-1)^i c^*_{ij} u_{ij}(X) \right\} \right] + \sum_{j\in[p]} c^*_{1j}. \end{aligned}$$

□

A.4 Proof for Corollary 7

Proof. Assume that the Rawls classifier f^* is non-trivial. By Definition 3 and the Rawls classifier f^* characterized by Theorem 6, the Rawls error rate is equal to

$$r^* = \max_{i\in\{0,1\}, j\in[p]} r_{ij}(f^*) = \sum_{i\in\{0,1\}, j\in[p]} c^*_{ij} r_{ij}(f^*).$$

Thus, the support of the convex combination given by c^* must be only over indices ij's where $r_{ij}(f^*)$ attains that maximum value. Equivalently, $\{ij : c^*_{ij} > 0\}$ must be the same as $\{ij : r_{ij}(f^*) = r^*\}$. Let ij be the unique index such that $c^*_{ij} > 0$ and let $c^*_{kl} = 0$ for all $(k,l) \neq (i,j)$. Then

$$f^*(x) = \begin{cases} \mathbb{1}\left(-c^*_{0j} u_{0j}(x) \geq 0\right) \equiv 0, & \text{if } i = 0, \\ \mathbb{1}\left(c^*_{1j} u_{1j}(x) \geq 0\right) \equiv 1, & \text{if } i = 1, \end{cases}$$

because $u_{ij}(x) = \Pr(Y = i, Z = j \mid X = x) / \Pr(Y = i, Z = j) \geq 0$, for all $i \in \{0,1\}, j \in [p]$ and $x \in \mathcal{X}$. This contradicts the non-triviality of f^*. □

A.5 Proof for Corollary 8

Proof. Assume that the Rawls classifier is non-trivial. Suppose there are no $j, k \in [p]$ such that S_{0j} and S_{1k} attain the Rawls rate r^*. Then the corresponding c^* that characterizes the Rawls classifier f^* in Theorem 6 must have either $\{ij : c^*_{ij} > 0\} \subseteq \{0j : j \in [p]\}$ or $\{1j : j \in [p]\}$. That is, the indices of the non-zero coordinates c^*_{ij} must either all have $i = 0$ or all have $i = 1$. In that case, the corresponding f^* by Theorem 6 looks like

$$f^*(x) = \begin{cases} \mathbb{1}\left(-\sum_{j\in[p]} c^*_{0j} u_{0j}(x) \geq 0\right) \equiv 0, & \text{if } i = 0, \\ \mathbb{1}\left(\sum_{j\in[p]} c^*_{1j} u_{1j}(x) \geq 0\right) \equiv 1, & \text{if } i = 1, \end{cases},$$

because $u_{ij}(x) = \Pr(Y = i, Z = j \mid X = x) / \Pr(Y = i, Z = j) \geq 0$, for all $i \in \{0,1\}, j \in [p]$ and $x \in \mathcal{X}$. This contradicts the non-triviality of f^*. □

A.6 Proof for Corollary 9

Proof. In the $p = 1$ case, the Rawls classifier f^* in Theorem 6 is given by some coefficients $\{c^*_{i1}\}_{i\in\{0,1\}}$, with $\sum_{ij} c^*_{ij} = 1$ and $c_{ij} \geq 0$, for all ij, such that

$$\begin{aligned} f^*(x) &= \mathbb{1}\left(c^*_{11} u_{11}(x) - c^*_{01} u_{01}(x) \geq 0\right) \\ &= \mathbb{1}\left(\frac{c^*_{11}}{p_{11}} \eta_{11}(x) - \frac{c^*_{01}}{p_{01}} \eta_{01}(x) \geq 0\right) \\ &= \mathbb{1}\left(\frac{c^*_{11}}{p_{11}} \eta_{11}(x) - \frac{c^*_{01}}{p_{01}} (1 - \eta_{11}(x)) \geq 0\right) \\ &= \mathbb{1}\left(\left(\frac{c^*_{11}}{p_{11}} + \frac{c^*_{01}}{p_{01}}\right) \eta_{11}(x) \geq \frac{c^*_{01}}{p_{01}}\right) \\ &= \mathbb{1}\left(\eta(x) \geq t\right), \end{aligned}$$

where $\eta(x) = \Pr(Y = 1 \mid X = x)$, and the threshold t is equal to

$$t = \frac{(c_{01}/p_{01})}{(c_{11}/p_{11}) + (c_{01}/p_{01})}.$$

□

B PROOF DETAILS FOR SECTION 4

B.1 Proof for Theorem 12

Proof. Since we are looking for classifiers $f_b(x) = \mathbb{1}\left(s(x) \geq b\right)$,

$$R_{ij}(f_b) = \sup_{(s,\mathcal{D})\in\mathcal{R}_{MV}} \Pr\left(f_b(s(X) \neq Y \mid Y = i, Z = j\right)$$

$$= \begin{cases} \sup_{(s,\mathcal{D})\in\mathcal{R}_{MV}} \Pr\left(s(X_{0j}) \geq b\right), & \text{for } i = 0 \\ \sup_{(s,\mathcal{D})\in\mathcal{R}_{MV}} \Pr\left(s(X_{1j}) \leq b\right), & \text{for } i = 1 \end{cases}$$

Therefore, the restricted Rawls error rate is given by the following optimization.

$$\text{minimize } \max_{j\in[p]} R_j \text{ over } b, R_1, R_2, \ldots, R_p \text{ subject to}$$

$$\sup_{(s,\mathcal{D})\in\mathcal{R}_{MV}} \Pr\left(s(X_{0j}) \geq b\right) \leq R_j,\ \forall j \in [p]$$

$$\sup_{(s,\mathcal{D})\in\mathcal{R}_{MV}} \Pr\left(s(X_{1j}) \leq b\right) \leq R_j,\ \forall j \in [p].$$

We can equivalently rewrite as

$$\text{maximize } \min_{j\in[p]} (1 - R_j) \text{ over } b, R_1, R_2, \ldots, R_p \text{ subject to}$$

$$\inf_{(s,\mathcal{D})\in\mathcal{R}_{MV}} \Pr\left(s(X_{0j}) \leq b\right) \geq 1 - R_j,\ \forall j \in [p]$$

$$\inf_{(s,\mathcal{D})\in\mathcal{R}_{MV}} \Pr\left(s(X_{1j}) \geq b\right) \geq 1 - R_j,\ \forall j \in [p].$$

Note that to satisfy $R_j \leq 1/2$, for all $j \in [p]$, we must have $\mu_{0j} \leq b \leq \mu_{1j}$, for all $j \in [p]$. Using the upper bound in the one-sided Chebyshev inequality (Proposition 14) and its tightness property, we can rewrite our optimization problem as follows.

$$\underset{b, R_1, R_2, \ldots, R_p}{\text{maximize}} \ \min_{j\in[p]} (1 - R_j) \quad \text{subject to}$$

$$\frac{(b - \mu_{0j})^2}{\sigma_{0j}^2 + (b - \mu_{0j})^2} \geq 1 - R_j,\ \forall j \in [p]$$

$$\frac{(\mu_{1j} - b)^2}{\sigma_{1j}^2 + (\mu_{1j} - b)^2} \geq 1 - R_j,\ \forall j \in [p].$$

Equivalently, the above can be written as

$$\underset{b, R_1, R_2, \ldots, R_p}{\text{maximize}} \ \min_{j\in[p]} (1 - R_j) \quad \text{subject to}$$

$$\frac{1}{1 - R_j} - 1 \geq \frac{\sigma_{0j}^2}{(b - \mu_{0j})^2},\ \forall j \in [p]$$

$$\frac{1}{1 - R_j} - 1 \geq \frac{\sigma_{1j}^2}{(\mu_{1j} - b)^2},\ \forall j \in [p].$$

Further simplification gives

$$\underset{b, R_1, R_2, \ldots, R_p}{\text{maximize}} \ \min_{j\in[p]} (1 - R_j) \quad \text{subject to}$$

$$b \geq \mu_{0j} + \sigma_{0j}\sqrt{\frac{1 - R_j}{R_j}},\ \forall j \in [p]$$

$$\mu_{1j} - \sigma_{1j}\sqrt{\frac{1 - R_j}{R_j}} \geq b,\ \forall j \in [p].$$

Note that the constraints arising from negative square root hold vacuously because $\mu_{0j} \leq b \leq \mu_{1j}$, for all $j \in [p]$. Now we can eliminate b from this optimization as follows.

$$\underset{R_1, R_2, \ldots, R_p}{\text{maximize}} \ \min_{j\in[p]} (1 - R_j) \quad \text{subject to}$$

$$\mu_{1j} - \sigma_{1j}\sqrt{\frac{1 - R_j}{R_j}} \geq \mu_{0j} + \sigma_{0j}\sqrt{\frac{1 - R_j}{R_j}},\ \forall j \in [p].$$

Equivalently, this can be written as

$$\underset{R_1, R_2, \ldots, R_p}{\text{maximize}} \ \min_{j\in[p]} (1 - R_j) \quad \text{subject to}$$

$$\frac{\mu_{1j} - \mu_{0j}}{\sigma_{1j} + \sigma_{0j}} \geq \sqrt{\frac{1 - R_j}{R_j}},\ \forall j \in [p].$$

Since $\sqrt{(1 - a)/a}$ is a monotonically decreasing function for $a > 0$, we need to only find

$$j^* = \underset{j\in[p]}{\text{argmin}} \ \frac{\mu_{1j} - \mu_{0j}}{\sigma_{1j} + \sigma_{0j}}.$$

Thus, the optimal η_{j^*} is

$$\eta_{j^*} = \left(1 + \left(\frac{\sigma_{1j^*} + \sigma_{0j^*}}{\mu_{1j^*} - \mu_{0j^*}}\right)^2\right)^{-1},$$

and the optimal threshold b^* is

$$b^* = \mu_{1j^*} - \sigma_{1j^*}\sqrt{\frac{\mu_{1j^*} - \mu_{0j^*}}{\sigma_{1j^*} + \sigma_{0j^*}}}$$

$$= \mu_{0j^*} + \sigma_{0j^*}\sqrt{\frac{\mu_{1j^*} - \mu_{0j^*}}{\sigma_{1j^*} + \sigma_{0j^*}}}.$$

□

Here we state the one-sided Chebyshev inequality for completeness. Its bound on the deviation away from the mean of a random variable plays a key role in our proof.

Proposition 14. *For any real-valued random variable X with mean μ and standard deviation σ, and for any $a > 0$,*

$$\Pr(X \geq \mu + a) \leq \frac{\sigma^2}{\sigma^2 + a^2}$$

$$\Pr(X \leq \mu - a) \leq \frac{\sigma^2}{\sigma^2 + a^2}.$$

Moreover, for any given $a > 0$, there exists a distribution X with mean μ and standard deviation σ such that these inequalities are tight.

B.2 Proof for Theorem 13

Proof. Let $s(X_{ij})$ be a Gaussian with mean $\mu_{ij} \in \mathbb{R}^d$ and covariance matrix $\Sigma_{ij} \in \mathbb{R}^{d\times d}$, respectively, for $i \in \{0, 1\}$ and $j \in [p]$. Let $f_{w,b}$ denote a linear classifier given by $f_{w,b}(x) = \mathbb{1}\left(w^T s(x) \geq b\right)$, for some $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$. Then, for $i \in \{0, 1\}$ and $j \in [p]$, we get

$$\Pr\left(f_{w,b}(X_{ij}) = 0\right) = \Pr\left(w^T s(X_{ij}) \leq b\right)$$

$$= \Phi\left(\frac{b - w^T \mu_{ij}}{\left\|\Sigma_{ij}^{1/2} w\right\|_2}\right),$$

where $\Phi(\cdot)$ is the CDF of a standard normal variable $N(0,1)$. Similarly, using $\Phi(t)+\Phi(-t)=1$, we have

$$\begin{aligned}\Pr\left(f_{w,b}(X_{ij})=1\right) &= \Pr\left(w^T s(X_{ij})\geq b\right)\\ &= \Phi\left(\frac{w^T\mu_{ij}-b}{\left\|\Sigma_{ij}^{1/2}w\right\|_2}\right).\end{aligned}$$

Thus, assuming that $s(X_{ij})$'s are Gaussians with the given means and covariance matrices, the restricted Rawls error rate can be obtained by solving the following optimization problem (where error rate constraints are replaced by accuracy constraints).

$$\begin{aligned}&\underset{w,b,r_1,r_2,\ldots,r_p}{\text{maximize}}\ \min_j(1-r_j)\quad\text{subject to}\\ &\Phi\left(\frac{b-w^T\mu_{0j}}{\left\|\Sigma_{0j}^{1/2}w\right\|_2}\right)\geq 1-r_j\quad\text{and}\\ &\Phi\left(\frac{w^T\mu_{1j}-b}{\left\|\Sigma_{1j}^{1/2}w\right\|_2}\right)\geq 1-r_j,\ \forall j\in[p].\end{aligned}$$

In other words, $\underset{w,b,r_1,r_2,\ldots,r_p}{\text{maximize}}\ \min_j(1-r_j)$ subject to

$$\begin{aligned}&b\geq w^T\mu_{0j}+\Phi^{-1}(1-r_j)\ \left\|\Sigma_{0j}^{1/2}w\right\|_2\quad\text{and}\\ &w^T\mu_{1j}-\Phi^{-1}(1-r_j)\ \left\|\Sigma_{1j}^{1/2}w\right\|_2\geq b,\ \forall j\in[p].\end{aligned}$$

Observe that b can be eliminated in the above to get

$$\begin{aligned}&\underset{w,r_1,r_2,\ldots,r_p}{\text{maximize}}\ \min_j(1-r_j)\quad\text{subject to }\forall j\in[p],\\ &w^T\mu_{0j}+\Phi^{-1}(1-r_j)\ \left\|\Sigma_{0j}^{1/2}w\right\|_2\\ &\leq w^T\mu_{1j}-\Phi^{-1}(1-r_j)\ \left\|\Sigma_{1j}^{1/2}w\right\|_2,\end{aligned}$$

Moreover, at the optimum $(w_*,b_*,r_1^*,r_2^*,\ldots,r_p^*)$, the constraint must be tight for some $j^*\in[p]$, giving

$$\begin{aligned}b_* &= w_*^T\mu_{0j^*}+\Phi^{-1}(1-r_{j^*})\ \left\|\Sigma_{0j^*}^{1/2}w_*\right\|_2\\ &= w_*^T\mu_{1j^*}-\Phi^{-1}(1-r_j^*)\ \left\|\Sigma_{1j^*}^{1/2}w_*\right\|_2.\end{aligned}$$

Using change of variables $\kappa_j=\Phi^{-1}(1-r_j)$ and monotonicity of $\Phi(\cdot)$, we can rewrite our optimization as follows.

$$\begin{aligned}&\underset{w,\kappa_1,\kappa_2,\ldots,\kappa_p}{\text{maximize}}\ \min_j\kappa_j\quad\text{subject to}\\ &w^T(\mu_{1j}-\mu_{0j})\geq\kappa_j\left(\left\|\Sigma_{1j}^{1/2}w\right\|_2+\left\|\Sigma_{0j}^{1/2}w\right\|_2\right),\ \forall j\in[p].\end{aligned}$$

We can eliminate κ_j's from this to write it as an equivalent optimization problem only over w.

$$\begin{aligned}&\underset{w}{\text{maximize}}\ \min_j\frac{w^T(\mu_{1j}-\mu_{0j})}{\left\|\Sigma_{1j}^{1/2}w\right\|_2+\left\|\Sigma_{0j}^{1/2}w\right\|_2},\ \text{equivalently,}\\ &\underset{w}{\text{minimize}}\ \max_j\frac{\left\|\Sigma_{1j}^{1/2}w\right\|_2+\left\|\Sigma_{0j}^{1/2}w\right\|_2}{w^T(\mu_{1j}-\mu_{0j})}.\end{aligned}$$

If $\mu_{1j}=\mu_{0j}$, for any $j\in[p]$, then the optimum κ^* must be zero. Otherwise, observe that the above constraints are homogeneous in w, i.e., whenever w satisfies these constraints, any non-negative scalar multiple of w also satisfies them. It is known that the optimization problem mentioned above can be solved efficiently in polynomial time using SOCP (Second Order Cone Programming) [4] and fractional programming [37]. A general framework to solve such problem is discussed in [30], which we defer to the full version of this paper. For now, it can be seen that when the covariance matrices are diagonal, this problem can be solved using SOCP.

We can assume $p\leq d$, that is the number of sensitive groups (e.g., race, gender) is smaller than the dimensionality of the feature space. In this case, WLOG, we can rewrite the above optimization as

$$\begin{aligned}&\underset{w}{\text{minimize}}\ \max_j\left\|\Sigma_{1j}^{1/2}w\right\|_2+\left\|\Sigma_{0j}^{1/2}w\right\|_2\quad\text{subject to}\\ &w^T(\mu_{1j}-\mu_{0j})=1,\ \forall j\in[p].\end{aligned}$$

There is a nice geometric interpretation for this optimal solution in the case of spherical Gaussians, i.e., when each covariance matrix $\Sigma_{ij}=\sigma_{ij}^2I$, for some $\sigma_{ij}\geq 0$ and the identity matrix $I\in\mathbb{R}^{d\times d}$. Then the above optimization becomes

$$\underset{w}{\text{minimize}}\ \max_j\frac{\|w\|_2\,(\sigma_{1j}+\sigma_{0j})}{w^T(\mu_{1j}-\mu_{0j})}.$$

Or equivalently,

$$\begin{aligned}&\underset{w}{\text{minimize}}\ \|w\|_2\ \text{subject to}\\ &w^T(\mu_{1j}-\mu_{0j})\geq\sigma_{1j}+\sigma_{0j},\ \forall j\in[p].\end{aligned}$$

Geometrically, this corresponds to the following classical problem [40]: given a polyhedron, find a point on it that is closest to the origin in ℓ_2 norm. This is a special case of convex quadratic optimization and is known to be solvable efficiently in polynomial time using the Ellipsoid method as well as the interior point method [10], and its complexity is similar to that of the support vector machine problem. □