Fairness without Harm: Decoupled Classifiers with Preference Guarantees

Berk Ustun [1] **Yang Liu** [2] **David C. Parkes** [1]

Abstract

In domains such as medicine, it can be acceptable for machine learning models to include *sensitive attributes* such as gender and ethnicity. In this work, we argue that when there is this kind of *treatment disparity* then it should be in the best interest of each group. Drawing on ethical principles such as beneficence ("do the best") and non-maleficence ("do no harm"), we show how to use sensitive attributes to train decoupled classifiers that satisfy *preference guarantees*. These guarantees ensure the majority of individuals in each group prefer their assigned classifier to (i) a pooled model that ignores group membership (*rationality*), and (ii) the model assigned to any other group (*envy-freeness*). We introduce a recursive procedure that adaptively selects group attributes for decoupling, and present formal conditions to ensure preference guarantees in terms of generalization error. We validate the effectiveness of the procedure on real-world datasets, showing that it improves accuracy without violating preference guarantees on test data.

1. Introduction

When machine learning systems are deployed in human-facing applications (e.g., lending, hiring, medical decision support), their performance may vary over groups defined by *sensitive attributes* such as gender and ethnicity. Such performance disparities are now regularly reported (Angwin et al., 2016; Dastin, 2018), eliciting calls for fairness in machine learning (Crawford, 2013), and prompting the development of technical solutions (Zliobaite, 2015; Barocas et al., 2018; Corbett-Davies & Goel, 2018).

[1]Harvard University, Cambridge, MA, USA [2]UC Santa Cruz, Santa Cruz, CA, USA. Correspondence to: Berk Ustun <berk@seas.harvard.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Many of the proposed methods for fair machine learning have aimed to build models that predict or perform in the same way across groups (e.g., Hardt et al., 2016; Zafar et al., 2017a; Feldman et al., 2015; Zafar et al., 2017c; Agarwal et al., 2018; Narasimhan, 2018). Such methods can be broadly viewed as methods to achieve fairness by *parity* (see Zafar et al., 2017b, for a discussion). Parity is an appropriate notion of fairness for applications such as hiring or sentencing, where a model that exhibits disparate treatment or disparate impact may be viewed as a system to perpetrate wrongful discrimination (see Arneson, 2006; Hellman, 2008; Barocas & Selbst, 2016).

In comparison, less work has sought to articulate suitable notions of fairness for domains with different ethical principles (with some exceptions, see e.g., Chen et al., 2018). In medical applications, for example, the relevant ethical principles are *beneficence* (do the best in one's ability) and *non-maleficence* (do no harm) (see e.g., Beauchamp et al., 2001). Accordingly, methods for fair machine learning should be designed to produce the most accurate model for each group (beneficence) without harming any group (non-maleficence).

These goals represent new challenges for the fair use of sensitive attributes in machine learning. Consider, for example, training a medical diagnostic using a dataset with sensitive attributes such as age, gender and ethnicity. In this case, a model that ignores group membership may not be beneficent as it may impose inevitable performance trade-offs between heterogeneous groups (see Figure 1). In practice, heterogeneity may arise due to intrinsic differences between groups, or discrepancies in the quality or amount of data.

While these issues motivate the need to build models that explicitly consider group membership (see Corbett-Davies et al., 2017; Lipton et al., 2018), it is not clear *how to do this in a way that is fair to each group*. As shown in Figure 2, simple approaches such as a "one-hot encoding" may not recover the most accurate model for each group. Conversely, one could harm groups by fitting a model from a hypothesis class that is overly complex (e.g., by overfitting), or that one that cannot adequately capture the heterogeneity (e.g., by "gerrymandering" along intersectional subgroups as discussed in Kearns et al., 2018; Hébert-Johnson et al., 2018).

	Group A			Group B			Pooled		
(x_1, x_2)	n^+	n^-	h^*_A	n^+	n^-	h^*_B	n^+	n^-	$\hat{h}_0$
(0, 0)	50	101	-	100	50	+	150	151	-
(0, 1)	101	50	+	50	100	-	151	150	+
(1, 0)	101	50	+	50	100	-	151	150	+
(1, 1)	101	50	+	50	100	-	151	150	+

Figure 1. Training a pooled classifier that ignores group membership may impose unavoidable trade-offs between groups. We are given data from two groups $z \in \{A, B\}$ with heterogeneous data distributions $\mathbb{P}(y = +1|\boldsymbol{x}, A) = \mathbb{P}(y = -1|\boldsymbol{x}, B)$. Here, n^+ and n^- denote the number of training examples with $y = +1$ and $y = -1$. Decoupled training produces the best classifier for each group $\hat{h}_A = h^*_A$ and $\hat{h}_B = h^*_B$, both of which have an error rate of 33%. In contrast, pooled training produces a classifier $\hat{h}_0$ with disparate impact due to a *tyranny of the majority*: the data contains slightly more samples from A so that empirical risk minimization outputs the *best* classifier for A which is the *worst* classifier for B. Pooled training with a parity constraint such as equal accuracy between A and B would fix the performance gap, but achieve an error rate of 50% for each group, missing the opportunity to provide better accuracy.

Group A				Group B				Pooled with z			
x_1	n^+	n^-	h^*_A	x_1	n^+	n^-	h^*_B	(x_1, z)	n^+	n^-	h^*_0
0	50	0	-	0	0	50	+	(0,0)	0	50	+
1	0	50	+	1	50	0	-	(1,0)	50	0	-
								(0,1)	50	0	-
								(1,1)	0	50	+

Figure 2. A pooled classifier that encodes group membership may not perform as well as a pair of decoupled classifier when we fit classifiers from a hypothesis class that cannot represent the heterogeneity between groups. Here, we consider training linear classifiers using data from heterogeneous groups $z \in \{A, B\}$. A linear classifier trained separately for each group has zero error. However, there does not exist a linear, pooled classifier with zero error due to the XOR structure.

In this paper, we aim to use sensitive attributes in a way that is aligned with the principles of beneficence and non-maleficence. Towards beneficence, we make use of decoupled classifiers— i.e., train a classifier for each group using data from that group. Decoupling is a simple technique that will recover the most accurate model for each group in an ideal setting where we are given unlimited data. In practice, however, it must be used with care since it may harm groups with insufficient data. Towards non-maleficence, we adopt the use of *preference guarantees*, which are a variation on those suggested by Zafar et al. (2017b). We require that each group should prefer their assigned model to (i) a pooled model that ignores group membership (*rationality*) and (ii) the model assigned to any other group (*envy-freeness*).

In settings where individuals prefer more accurate models, rationality and envy-freeness ensure that the majority of individuals in each group would choose to report their sensitive attributes if they were allowed to not report them (thus opting for a pooled model) or to misreport them (thus opting for the model assigned to another group).

The main contributions of this paper are:

- We present formal conditions for fair decoupling, i.e., that the preference guarantees of rationality and envy-freeness are satisfied. This is non-trivial because we require these properties to hold with respect to generalization error.
- We develop a recursive partitioning procedure to train decoupled classifiers for groups specified by multiple sensitive attributes without violating their preferences.
- We pair our procedure with an integer programming method to train linear classifiers via 0-1 loss minimization. This produces classifiers that satisfy preferences on training data and avoids pitfalls of surrogate loss minimization in our setting.
- We present experiments on real-world datasets that show that our procedure can output classifiers with good accuracy and that are responsive to preference guarantees.

Related Work We build on the work of Zafar et al. (2017b), who present the preference-based notions of *envy-freeness*, as well as *preferred impact*, which requires that groups prefer their assigned classifier to a pooled classifier trained with a parity constraint. Their method trains a linear classifier for each group by solving a *coupled* empirical risk minimization problem that enforces preferences with a convex surrogate loss function. In our experiments, we show that this approach may violate preference guarantees, even on training data. In contrast, decoupled training via 0-1 loss minimization immediately achieves preference guarantees on training data (see Remark 1), and is developed here as an adaptive procedure that ensures preferences on test data.

Our paper also builds on the work of Dwork et al. (2018), who study decoupling as a way to achieve parity-based notions of fairness. Their work presents impossibility results to motivate decoupling (i.e., a bound on the maximum loss of accuracy due to pooled training for different hypothesis classes), as well as a computationally efficient procedure to optimize a joint loss function on a set of classifiers (see also Alabi et al., 2018). While their work does not consider preference guarantees, it does caution that decoupling may harm groups with insufficient data. They propose mitigating the harm by transfer learning, but their experiments show that this approach may still result in harm.

Our work is broadly related to several streams in the literature on fair machine learning. Our goals resemble those of Chen et al. (2018), who present a beneficent approach to reduce performance disparities between groups via data collection. Our method aims to ensure preferences in terms of generalization error (c.f. Woodworth et al., 2017; Cotter et al., 2018) and among intersectional groups (c.f., Kearns

et al., 2018; Hébert-Johnson et al., 2018). We allow data distributions to vary between groups, sharing this motivation with causal approaches (see e.g., Kusner et al., 2017; Nabi & Shpitser, 2018; Zhang & Bareinboim, 2018; Salimi et al., 2019). In addition to Zafar et al. (2017b) and Dwork et al. (2018), several other works discuss treatment disparity to achieve parity-based notions of fairness (see e.g., Corbett-Davies & Goel, 2018; Kleinberg et al., 2018; Lipton et al., 2018; Wang et al., 2019).

2. Problem Statement

We start with a dataset with n examples $(\boldsymbol{x}_i, y_i, z_i)_{i=1}^n$, where each example consists of a *feature vector* $\boldsymbol{x}_i = [1, x_{i,1}, \ldots, x_{i,d}] \in \mathbb{R}^{d+1}$, a *label* $y_i \in \{\pm 1\}$, and a vector of m *group attributes* $z_i = [z_{i,1}, \ldots, z_{i,m}] \in Z$ (e.g., $z_i = [\texttt{female}, \texttt{old}]$). We denote the indices of examples in group z as $I_z = \{i \,|z_i = z\}$, and let $n_z = |I_z|$. Since group attributes are categorical, Z partitions the data so that $\cup_{z \in Z} I_z = \{1, \ldots, n\}$ and $I_z \cap I_{z'} = \emptyset$ for all $z, z' \in Z$.

We use the dataset to train a set of classifiers for each group, which we denote as $H_Z = \{\hat{h}_z\}_{z \in Z}$. We assume that all classifiers belong to the same hypothesis class $h_z \in \mathcal{H}$ for all $z \in Z$. Given a classifier $h : \mathbb{R}^{d+1} \to \{\pm 1\}$, we denote its *empirical risk* (i.e., training error) and *true risk* (i.e., generalization error) for group z as

$$\hat{R}_z(h) = \frac{1}{n_z} \sum_{i \in I_z} \mathbf{1}_{y_i \neq h(\boldsymbol{x}_i)}, \quad R_z(h) = \mathbb{E}_{\boldsymbol{x}, y|z} \left[\mathbf{1}_{y \neq h(\boldsymbol{x})} \right].$$

We define $\hat{h}_z = \operatorname{argmin}_{h \in \mathcal{H}} \hat{R}_z(h)$.

Preference Guarantees In an ideal setting where we are given sufficient data from each group, decoupling would recover the *most accurate classifier for each group*: $h_z^* = \operatorname{argmin}_{h \in \mathcal{H}} R_z(h)$. In practice, decoupling cannot be expected to recover h_z^* for each group given the lack of training data. Nevertheless, it may uniformly improve the performance for all groups with respect to a *pooled classifier* $\hat{h}_0 = \operatorname{argmin}_{h \in \mathcal{H}} \hat{R}(h)$ trained without sensitive attributes.

We stipulate that a set of decoupled classifiers $H_Z = \{\hat{h}_z\}_{z \in Z}$ should be deployed over a pooled classifier $\hat{h}_0$ when they provide the following guarantees:

Rationality. A set of decoupled classifiers satisfies rationality if each group is assigned a model that is at least as accurate as the pooled classifier: $R_z(\hat{h}_z) \leq R_z(\hat{h}_0)$ for all $z \in Z$.

Envy-freeness. A set of decoupled classifiers satisfies envy-freeness if each group is assigned a classifier that is at least as accurate as the classifiers assigned to other groups: $R_z(\hat{h}_z) \leq R_z(\hat{h}_{z'})$ for all $z, z' \in Z$.

Note that both guarantees are defined in terms of *true risk*. Rationality and envy-freeness follow without loss of generality when we recover the best model for each group. These conditions also enshrine basic principles of fairness for how we should use sensitive attributes in prediction. Specifically, these guarantees ensure that the majority in each group would choose their assigned model given a preference for low generalization error. Without rationality, a majority in some group would prefer the pooled model. Without envy-freeness, a majority in some group would prefer the model assigned to another group.

We evaluate the preference of group z between a pair of classifiers, h and h', using the *preference gap measures*:

$$\hat{\Delta}_z(h, h') = \hat{R}_z(h) - \hat{R}_z(h') \tag{1}$$

$$\Delta_z(h, h') = R_z(h) - R_z(h'), \tag{2}$$

We can measure preference gaps in terms of empirical risk (1), but care about the preference gap in terms of true risk (2). Thus, a set of decoupled classifiers $H_Z = \{\hat{h}_z\}_{z \in Z}$ satisfies rationality with respect to a pooled model $\hat{h}_0$ if $\Delta_z(\hat{h}_z, \hat{h}_0) \geq 0$ for all $z \in Z$, and satisfies envy-freeness if $\Delta_z(\hat{h}_z, \hat{h}_{z'}) \geq 0$ for all $z, z' \in Z$.

Choosing Attributes for Decoupling Consider a case where we train classifiers for groups specified by age and gender. Ideally, we would train decoupled classifiers for the most granular groups. However, this may not be feasible nor safe: we may have no data for some groups (e.g., young females), or we may train classifiers that violate rationality or envy-freeness. In such cases, it may be possible to improve accuracy for all groups by decoupling along a carefully chosen subset of group attributes.

We formalize this problem as follows. Given a set of m *atomic* group attributes $Z = Z_1 \times \ldots \times Z_m$, we allow for decoupling along a subset of group attributes V. We represent the assignment of decoupled classifiers to groups as a *tree* $T = (V_T, H_T)$, where V_T denotes a set of groups and $H_T = \{\hat{h}_v\}_{v \in V_T}$ denotes the set of classifiers train using the data for each $v \in V_T$ (see Figure 3). Given a tree T, we denote the assignment of classifiers from V_T to the groups in Z with an *assignment function* $a(\cdot) : Z \to T$. Thus, $\hat{h}_{a(z)}$ is the classifier assigned to group z by tree T.

We aim to find a set of decoupled classifiers that respects a tree structure by solving the optimization problem:

$$\begin{aligned} \min_{T \in \mathcal{T}(Z)} \quad & \Phi(T) \\ \text{s.t.} \quad & \Delta_z(\hat{h}_{a(z)}, \hat{h}_0) \geq 0 \quad \forall z \in Z, \quad (3a) \\ & \Delta_z(\hat{h}_{a(z)}, \hat{h}_{a(z')}) \geq 0 \quad \forall z, z' \in Z. \quad (3b) \end{aligned}$$

Here, $\mathcal{T}(Z)$ is the set of all trees – i.e., all possible ways to assign classifiers to groups specified by Z. Constraints (3a) and (3b) require a set of decoupled classifiers to satisfy rationality and envy-freeness. $\Phi(\cdot)$ is a *cost function* to

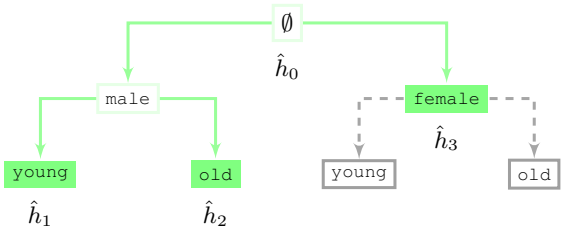

Figure 3. A set of decoupled classifiers assigned to 4 groups defined by 2 sensitive attributes $Z = (\texttt{male}, \texttt{female}) \times (\texttt{young}, \texttt{old})$. Here, we train the classifiers $H_T = \{\hat{h}_1, \hat{h}_2, \hat{h}_3\}$ using the data at the leaves $V_T = \{(\texttt{young}, \texttt{male}), (\texttt{old}, \texttt{male}), (\texttt{female})\}$. The tree structure ensures that decoupled classifiers are trained using the data pertaining to groups with shared sensitive attributes.

choose between a set of decoupled classifiers with preference guarantees.

Our decoupling procedure can handle any cost function, and will strive to optimize cost only after it has found a tree that does not violate preferences. Illustrative cost functions include:

- *Worst-Case Group Risk*: $\max_{z \in Z} R_z(\hat{h}_{a(z)})$, which reflects the worst error incurred by any group that is assigned to its own classifier (see e.g., Hashimoto et al., 2018).
- *Population Risk*: $\sum_{z \in Z} \pi_z R_z(\hat{h}_{a(z)})$, which reflects the aggregate generalization error over a population of interest. Here, π_z is the probability that an individual belongs to group z. These weights can be set as $\pi_z = n_z/n$ by default, or used to correct for systematic sampling bias.

Atomic Groups vs. Assigned Groups Our definitions of rationality and envy-freeness apply to the most granular groups that can be specified by sensitive attributes – i.e., for the atomic groups $z \in Z$. This reflects a notion that is robust against the possibility of "gerrymandering" along sensitive attributes (see e.g., Kearns et al., 2018; Hébert-Johnson et al., 2018). In settings where these guarantees are too strong given the available data and the number of atomic groups, one could also consider relaxing the definitions so that preferences hold for only for the groups generated by the decoupling procedure i.e., for each $v \in V_T$. Although our decoupling procedure can handle both settings, we adopt the stronger definitions throughout our paper.

3. Preference Guarantees

In this section, we present formal conditions for decoupled classifiers to satisfy preference guarantees.

We first observe that decoupled classifiers will satisfy rationality and envy-freeness on training data if we directly minimize the error rate (i.e., via the 0-1 loss function).

Remark 1 *A set of decoupled classifiers $H_Z = \{\hat{h}_z\}_{z \in Z}$, will satisfy rationality and envy-freeness on training data*

$$\hat{\Delta}_z(\hat{h}_z, \hat{h}_0) \geq 0 \quad \text{and} \quad \hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'}) \geq 0$$

for all $z, z' \in Z$ so long as $\hat{h}_z \in \operatorname{argmin}_{h \in \mathcal{H}} \hat{R}_z(h)$ for each $z \in Z$.

As shown in Figure 4, classifiers trained with a surrogate loss function (as in Zafar et al., 2017b) do not provide such guarantees, and may not satisfy rationality and envy-freeness on training data. Such violations can stem from a lack of data for some groups, or the fact that surrogate losses may not be robust to outliers (e.g., points belonging to a heterogeneous subpopulation; Brooks, 2011; Nguyen & Sanner, 2013).

In Theorem 2, we present a sufficient condition for a set of decoupled classifiers to satisfy rationality and envy-freeness (see Appendix A for a proof).

Theorem 2 *Given a set of decoupled classifiers $H_Z = \{\hat{h}_z\}_{z \in Z}$, denote the minimal empirical preference gap of group z as*

$$\hat{\epsilon}_z = \min\left(\hat{\Delta}_z(\hat{h}_z, \hat{h}_0), \min_{z' \in Z/\{z\}} \hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})\right).$$

Then H_Z satisfies rationality and envy-freeness with probability at least $1 - \delta$ so long as the following conditions hold for all groups $z \in Z$:

$$n_z \geq \frac{64 \ln |\mathcal{H}| + 4 \ln\left(\frac{2|Z|^2}{\delta}\right)}{\hat{\epsilon}_z^2} \qquad \text{and} \qquad \hat{\epsilon}_z > 0.$$

Theorem 2 has several implications for decoupled training in finite-sample settings:

- There exists a finite number of samples n_z after which decoupled classifiers satisfy rationality and envy-freeness.
- In finite-sample regimes, n_z is fixed. Thus, we obtain better guarantees by minimizing $|\mathcal{H}|$ while maximizing $\hat{\epsilon}_z$. This is why we train linear classifiers by optimizing the 0-1 loss.[1]
- In settings with a large number of groups (e.g. $|Z| \geq 10$), we require even more samples per group to provide preference guarantees. This may motivate the need to consider weaker preference guarantees that hold for the groups determined by our decoupling procedure (see Section 2).

[1]The bound in Theorem 2 is for a finite hypothesis class $\mathcal{H}$. However, an analogous bound can be derived by replacing $|\mathcal{H}|$ with a term based on the VC-dimension.

4. Training Decoupled Classifiers

In this section, we present a recursive procedure to train decoupled classifiers with preferences guarantees. Our procedure resembles decision tree methods in that it grows a tree with training data before pruning it with test data. In particular, our procedure:

- Uses a specialized routine to grow the tree, which generates a collection of "candidate" classifier sets that satisfy rationality and envy-freeness on the training data;
- Uses a specialized routine to prune the tree, which discards candidate classifier sets that violate rationality and envy-freeness in terms of generalization error.

We pair our procedure with an integer programming method to train linear classifiers by directly minimizing the 0-1 loss, which has several benefits in this setting.

4.1. Routine to Grow the Tree

We present the routine to grow our tree in Algorithm 4.1. Our routine induces a tree such as in Figure 3 by recursively spitting the training data, and training classifiers for the groups specified at each leaf node. Starting with the root node $\hat{h}_0$, which corresponds to $\{\emptyset\}$, each iteration aims to replace a classifier at a leaf node with a set of decoupled classifiers $\{\hat{h}_s\}_{s \in S}$ where $S \in \{Z_1, Z_2, \ldots, Z_m\}$.

In Algorithm 4.1, we describe this process as a search over *feasible splits*. Each split represents a distinct way to grow the tree, and is uniquely specified by a group attribute S that can be used for decoupling at a leaf node $v \in V_T$. Our routine considers all possible ways to grow the tree T. For each leaf $v \in V_T$, it calls the function $\mathsf{FeasibleSplits}(T, v, Z)$ to return all attributes S that: (i) have not already been used to decouple at v; and (ii) obey a user-specified sample size requirement.[2] For each $S \in \mathsf{FeasibleSplits}(T, v, Z)$, the routine trains a set of decoupled classifiers using the data at the leaf node v.

Once the routine has trained decoupled classifiers for all feasible splits, it chooses a split. For each S, it considers the tree $T_{v,S}$ produced by switching the pooled classifier at v with a set of classifiers for each $s \in S$. The routine assigns a score to each $T_{v,S}$ using the function $\mathsf{ViolationScore}(T_{v,S})$, which reflects the probability that the classifiers assigned by $T_{v,S}$ violate preferences. Formally, $\mathsf{ViolationScore}(T)$ is a bound on the probability that a classifier assigned by T violates rationality or envy-freeness (see Appendix B).

$$\mathbb{P}\left(\begin{smallmatrix}H_T \text{ violates}\\ \text{rationality or}\\ \text{envy-freeness}\end{smallmatrix}\right) \leq \mathsf{ViolationScore}(T)$$
$$= \sum_{z \in Z} 4 \exp\left(-\tfrac{n_z}{2} \cdot \hat{\Delta}_z(\hat{h}_{a(z)}, \hat{h}_0)^2\right) +$$
$$\sum_{z \in Z} \sum_{\substack{z' \in Z \\ a(z') \neq a(z)}} 4 \exp\left(-\tfrac{n_z}{2} \cdot \hat{\Delta}_z(\hat{h}_{a(z)}, \hat{h}_{a(z')})^2\right)$$

Algorithm 1 Recursive Decoupling

$T \leftarrow (\hat{h}_0, \{\emptyset\})$
repeat
  $\mathcal{T} \leftarrow []$
  for $v \in V_T$ **do**
    for $S \in \mathsf{FeasibleSplits}(T, v, Z)$ **do**
      $T_{v,S} \leftarrow \mathsf{Decouple}(T, v, S)$
      add $T_{v,S}$ to $\mathcal{T}$
    end for
  end for
  if $|\mathcal{T}| \geq 1$ **then**
    $T \leftarrow \text{argmin}_{T \in \mathcal{T}} \mathsf{ViolationScore}(T)$
  end if
until $\mathcal{T}$ is empty

procedure $\mathsf{Decouple}(T, v, S)$
replace the pooled classifier at v with a set of decoupled classifiers for each $s \in S$
  for $s \in S$ **do**
    $\hat{h}_s \leftarrow \text{argmin } \hat{R}_{v \wedge s}(h)$ ▷ fit model for $i \in I_v \cap I_s$
  end for
  $V' \leftarrow (V_T \setminus \{v\}) \cup S$
  $H' \leftarrow (H \setminus \{\hat{h}_v\}) \cup \{\hat{h}_s\}_{s \in S}$
  return $T' = (H', V')$
end procedure

4.2. Routine to Prune the Tree

We produce a collection of "candidate trees" by recording the tree at each iteration of Algorithm 4.1. We evaluate the rationality and envy-freeness of the classifiers assigned by each candidate tree using a hypothesis test on a hold-out dataset. We use an exact version of the *McNemar test*, which is commonly used to test for differences in the generalization error of classifiers (see e.g., Dietterich, 1998).

Given two classifiers h and h', we evaluate the preference of group z between h and h' for group z by testing:

$$H_0 : R_z(h) \leq R_z(h') \quad \text{vs.} \quad H_A : R_z(h) > R_z(h')$$

Here, the null hypothesis H_0 assumes that group z prefers h to h' by default. Thus, we reject H_0 when there is enough evidence to support a preference violation for group z in the hold-out data.

[2] For example, one may require any feasible S to satisfy the following conditions for each group $s \in S$: (i) contain at least 1 sample with each label; (ii) contain at least $n_s \geq d$ samples, where d is the number of variables. These are minimal conditions to ensure that we will not train a classifier that is linearly separable by default, or one that will trivially predict the majority class.

Given a tree with V_T leaves, we check the rationality and envy-freeness between groups using $|Z| + |Z|(|V_T| - 1)$ McNemar tests: $|Z|$ tests comparing the error of each group z between their assigned classifier $\hat{h}_{a(z)}$ and the pooled classifier; and $|Z|(|V_T| - 1)$ tests comparing the error of group z between their assigned classifier $\hat{h}_{a(z)}$ and any other classifier. We control the false discovery rate due to multiple testing using a standard Bonferroni correction (Dunn, 1961), which is suitable even for non-independent tests.[3]

We discard any tree that fails at least one test at a user-defined significance level. The remaining trees satisfy preferences on the hold-out data subject to a user-specified limit on type I error (i.e., a limit on the probability that the test incorrectly detects a preference violation that does not exist). Given the collection of remaining candidate trees, we then choose the tree that minimizes the cost function.

Our setup assumes rationality and envy-freeness by default, and only discards candidate trees if there is enough evidence to support a preference violation. In effect, this is an optimistic viewpoint, which we find justified given that the classifiers satisfy rationality and envy-freeness on the training data (since we will optimize the 0-1 loss).[4]

4.3. Direct Loss Minimization

Our procedure can be paired with any binary classification algorithm. Considering the results in Section 3, however, we pair it with a method to train linear classifiers by directly minimizing the 0-1 loss function. This has two important benefits in our setting:

- It produces a set of decoupled classifiers that are rational and envy-free on the training data (see Remark 1), which is not necessarily the case when we train classifiers by optimizing a surrogate loss function.
- Since 0-1 loss minimization satisfies rationality and envy-freeness on the training data, it ensures that our procedure will keep decoupling until it has grown a tree that assigns a classifier to each group $z \in Z$. This provides some protection against gerrymandering, in that the procedure will always consider assigning each group its own classifier, and only assign a classifier to multiple groups if this assignment violates preferences or does not optimize costs.

We train a linear classifier $h(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ that optimizes the 0-1 loss function by solving the MIP formulation:

$$
\begin{aligned}
\min \quad & \sum_{i=0}^{n} l_i \\
\text{s.t.} \quad & M_i l_i \geq y_i(\gamma - \sum_{j=0}^{d} w_j x_{ij}) && i = 1,...,n && \text{(4a)} \\
& 1 = l_i + l_{i'} && (i, i') \in K && \text{(4b)} \\
& w_j = w_j^+ + w_j^- && j = 0,...,d \\
& 1 = \sum_{j=0}^{d} (w_j^+ - w_j^-) && && \text{(4c)} \\
& l_i \in \{0, 1\} && i = 1,...,n \\
& w_j \in [-1, 1] && j = 0,...,d \\
& w_j^+ \in [0, 1] && j = 0,...,d \\
& w_j^- \in [-1, 0] && j = 0,...,d
\end{aligned}
$$

Here, constraints (4a) set the mistake indicators $l_i \leftarrow 1[h(\boldsymbol{x}_i) \neq y_i]$. These constraints depend a margin parameter γ, which should be set to a small positive number (e.g., 10^{-4}), as well as "Big-M" parameters M_i, which can be bounded since we have fixed $\|\boldsymbol{w}\|_1 = 1$ in constraint (4c). Constraint (4b) produces an improved lower bound by encoding the necessary condition that any classifier must make exactly one mistake for two points $(i, i') \in K$ with identical features $\boldsymbol{x}_i = \boldsymbol{x}_{i'}$ but conflicting labels. Here, $K = \{(i, i') : \boldsymbol{x}_i = \boldsymbol{x}_{i'}, y_i = +1, y_{i'} = -1\}$ is the set of points with conflicting labels.

While direct loss minimization is computationally intractable in the worst-case, it is often feasible through the use of modern integer programming tools (see e.g, Ustun & Rudin, 2016; Zeng et al., 2017). In our experiments in Section 5, we often train classifiers *to optimality* in minutes (see Table 1 for problem sizes). Algorithm 4.1 also provides several ways to reduce computation. For example, we can speed up the DECOUPLE procedure by using the pooled classifier to initialize training for each decoupled classifier (since the pooled classifier is a feasible solution to the MIP, this initializes the solver with an improved upper bound).

5. Experiments

We now present experiments that compare different methods to train classifiers with preference guarantees. We provide software to reproduce our results at https://www.github.com/ustunb/dcptree.

Setup We work with five datasets, each of which have multiple sensitive attributes, as shown in Table 1.[5] We pro-

[3] We cannot assume independence between the tests since we use the same hold-out set.

[4] Alternatively, one could use an inverted test where $H_0 : R_z(h') \geq R_z(h')$. This setup would reject H_0 only when there is sufficient evidence to support decoupling, which may be suitable for settings where, for example, we can assume that the data for each group is drawn from the same joint distribution.

[5] The datasets include: `adult`, the Adult dataset from the UCI ML Repository (Lichman, 2013); `arrest` and `violent`, the COMPAS recidivism dataset for arrest and violent crime (Angwin et al., 2016); `apnea`, a dataset to diagnose obstructive sleep apnea (Ustun et al., 2016); and `cancer`, a dataset to diagnose lung cancer (National Lung Screening Trial Research Team, 2011).

cess each dataset to define groups with a minimum number of samples and repair class imbalances at the group level.

We train linear classifiers using the following methods:

- LR: Decoupled logistic regression.
- DCCP: The coupled convex-concave programming approach of Zafar et al. (2017b).
- TREE01 & TREELR: Our procedure paired with 0-1 loss minimization and logistic regression, respectively. We allocate a third of the training data to the pruning procedure, and discard trees that violate rationality or envy-freeness at a significance level of 10%. The final tree minimizes the worst-case group risk (see Section 2).

We evaluate the performance and preferences of all classifiers for the atomic groups Z on a test set containing 25% of examples. Since LR and DCCP cannot decouple adaptively along sensitive attributes, we test them in two regimes: (i) train separate models for groups defined by *one attribute*; (ii) train separate models for groups defined by *all attributes*. Since there are multiple attributes that can be used in regime (i), we train classifiers for each attribute, and show results for the one that optimizes overall test accuracy in Table 1.

		ONE ATTRIBUTE		ALL ATTRIBUTES		ADAPTIVE	
Dataset	**Metrics**	DCCP	LR	DCCP	LR	TREELR	TREE01
`adult` $m=12$ $n=49,440$ $d=28$	# violations	2	1	3	1	1	0
	max gain	2.9%	2.9%	12.6%	17.6%	4.1%	10.0%
	min envy	2.9%	2.9%	26.1%	33.3%	11.8%	30.2%
	Δ disparity	0.8%	0.6%	0.3%	0.4%	-0.4%	1.9%
	# models	2	2	12	12	7	9
`apnea` $m=6$ $n=3,016$ $d=28$	# violations	0	0	0	0	0	0
	max gain	2.3%	1.2%	11.0%	13.7%	8.9%	30.8%
	min envy	5.4%	7.7%	21.2%	31.5%	21.2%	30.8%
	Δ disparity	-3.6%	-3.2%	7.1%	12.5%	7.6%	-6.8%
	# models	2	2	6	6	4	2
`arrest` $m=6$ $n=7,168$ $d=7$	# violations	0	0	2	3	1	0
	max gain	9.0%	7.7%	10.3%	11.5%	9.0%	2.2%
	min envy	9.0%	7.7%	14.1%	15.4%	12.8%	3.3%
	Δ disparity	-0.1%	-0.1%	-0.3%	1.0%	-2.7%	-2.6%
	# models	2	2	6	6	4	2
`cancer` $m=4$ $n=62,916$ $d=20$	# violations	1	2	1	1	1	0
	max gain	0.4%	0.4%	0.9%	1.0%	1.1%	2.8%
	min envy	3.4%	1.8%	2.1%	3.9%	3.8%	10.4%
	Δ disparity	1.0%	0.3%	-0.4%	0.0%	1.5%	0.8%
	# models	2	2	4	4	2	4
`violent` $m=6$ $n=10,960$ $d=7$	# violations	1	1	1	0	2	0
	max gain	9.3%	11.6%	13.6%	13.6%	9.3%	14.4%
	min envy	7.6%	10.2%	17.9%	17.0%	7.6%	22.0%
	Δ disparity	-9.7%	-4.1%	-13.6%	-14.0%	-9.8%	-8.4%
	# models	2	2	6	6	2	6

Table 1. Performance metrics for all methods on all datasets. We highlight methods that violate preferences in red and methods that satisfy preferences in green. Here, # models is the number of distinct classifiers assigned to the $m = |Z|$ atomic groups.

Results We present an overview of the performance for each method in Table 1. We report the following metrics:

- *# of violations*: the # of groups for which rationality is violated plus the # of pairwise comparisons between groups for which envy-freeness is violated. We evaluate each violation on the test data with a McNemar test at a 10% significance level.
- *max gain*: the maximum difference in accuracy between the pooled classifier and the decoupled classifier among all groups, $\max_{z \in Z} \Delta_z(\hat{h}_z, \hat{h}_0)$.
- *min envy*: the maximum difference in the degree of envy-freeness between groups, $\max_{z,z' \in Z} \Delta_z(\hat{h}_z, \hat{h}_{z'})$.
- Δ *disparity*: the disparity under the decoupled models minus the disparity under the pooled model, where disparity is measured as the maximum difference in the accuracy of any two groups.

We discuss key properties of our procedure with respect to the preference violations that can also be seen in Figure 4:

- Our approach produces classifiers that satisfy preference guarantees on training data and – almost always – test data. In particular, TREE01 achieves the smallest number of preference violations on test data across all datasets.
- When there exists a way to improve accuracy without harming groups, our approach tends to provide the largest possible improvement to each group (see e.g., `apnea`, decoupling has a max gain of 30.8%).
- The use of surrogate loss for the coupled training procedure of Zafar et al. (2017b) may produce classifiers that violate preferences, even on training data. Such violations can occur when groups contain too few training samples or when the data for one group contains outliers (e.g., data corresponding to a heterogeneous subpopulation).
- Decoupling along one sensitive attribute may lead to preference violations along smaller groups – highlighting the potential to achieve preference-based notions of fairness by "gerrymandering." Our approach has two benefits in this setting: (i) it provides a way to uniformly improve performance by assigning classifiers to "coarser" groups in a way that satisfies preference guarantees on atomic groups; (ii) it will always evaluate the feasibility of assigning each group its own classifier, and only resorts to the former option when there exists a better way that does not violate the preferences of atomic groups.
- Decoupling can occasionally lead to uniform improvements for all groups (see e.g., LR on `apnea`). We find that standard measures of error (e.g., aggregate test error) do not vary much across methods. However, there may be considerably larger changes for small groups. Our preference guarantees aim to ensure that decoupling benefits all groups without harming any group. As a result the disparities between groups may increase (or decrease), but in a way that is unlikely to lead to harm.

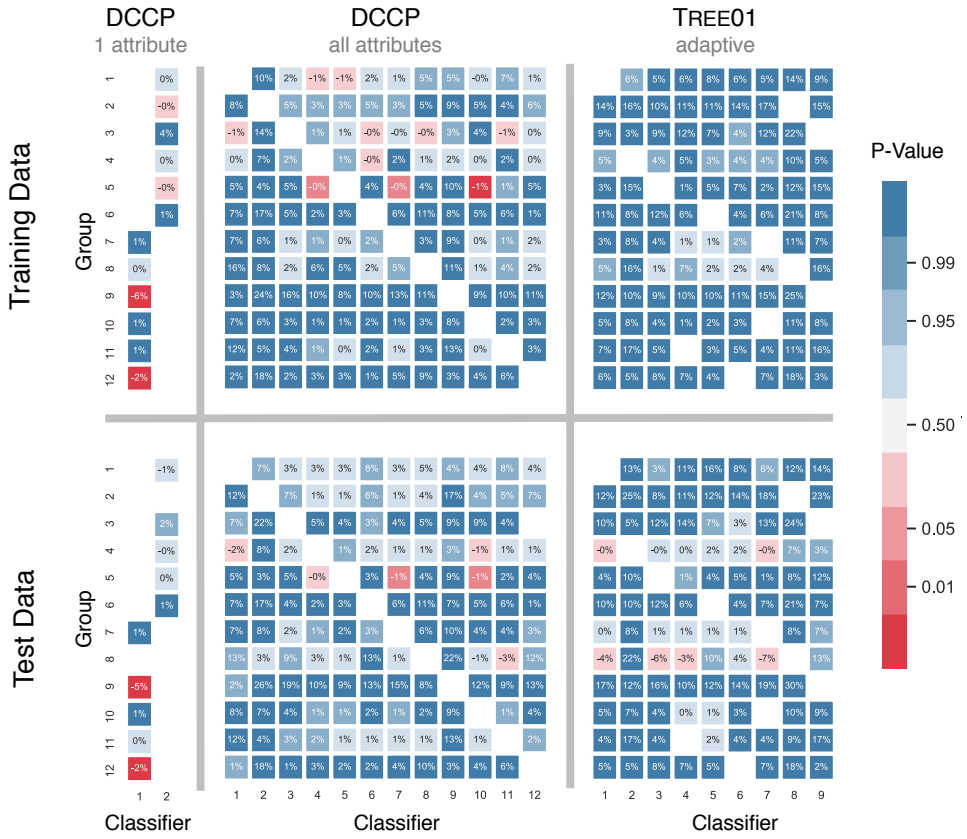

Figure 4. Envy-freeness gaps for decoupled classifiers trained on the `adult` dataset. The dataset contains $m = 12$ groups defined by 3 attributes (`gender`, `immigrant`, `marital_status`). We compare decoupled classifiers produced using our approach (right) to those built using DCCP. Since DCCP cannot decouple adaptively, we consider decoupling with a single binary attribute (`gender`) (left) and with all attributes (middle). For each method, we show how the accuracy for each group (y-axis) changes when they swap their assigned classifier with a classifier assigned to a different group (x-axis). We highlight cells based on the p-value of an envy-freeness violation, so that statistically significant violations appear in red. We observe: (i) our approach trains 9 decoupled classifiers that satisfy preferences for all 12 groups on training data and test data; (ii) DCCP with all attributes violates collective preferences on training data and test data; (iii) DCCP with one attribute leads to "gerrymandering" as envy-freeness is violated among the atomic groups.

6. Discussion

There are many domains where machine learning is used, sensitive attributes are readily collected, and it is legal to use them in prediction. In applications where groups benefit from improved accuracy, the principles of beneficence and non-maleficence suggest that we should aim to use sensitive attributes in a way that allows us to train the most accurate model for each group without harming any group.

We believe that attaining improved accuracy subject to these preference guarantees represents an important direction for future research in fair machine learning given growing calls for such methods in medicine (see e.g., Ferryman & Pitcan, 2018; Popejoy & Fullerton, 2016; Vayena et al., 2018; Chen et al., 2019). While parity-based methods are appropriate for some settings (e.g., risk adjustment formulas for healthcare spending as in Zink & Rose, 2019), they may be not be suitable for others due to their lack of beneficence (see e.g., Figure 1), and their potential to harm groups to achieve parity (see e.g., Lipton et al., 2018; Hu & Chen, 2019).

Our work highlights an important role for the preference guarantees of rationality and envy-freeness in such applications — i.e., as a set of formal criteria to ensure the fair use of sensitive attributes in prediction. In applications where treatment disparity may be legal, rationality and envy-freeness ensure that practitioners make use of sensitive attributes in a way that is aligned with the collective interests of each group. Specifically, they ensure that a majority of individuals in each group would opt for their assigned model rather than a model trained without sensitive attributes, or the model that would be assigned if they had changed or misreported group membership.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A Reductions Approach to Fair Classification. In *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2018.

Alabi, D., Immorlica, N., and Kalai, A. Unleashing linear optimizers for group-fair learning and optimization. In *Conference On Learning Theory*, pp. 2043–2066, 2018.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine Bias, 2016.

Arneson, R. J. What is wrongful discrimination? *San Diego L. Rev.*, 43:775, 2006.

Barocas, S. and Selbst, A. D. Big data's disparate impact. *Cal. L. Rev.*, 104:671, 2016.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning*. fairmlbook.org, 2018. http://www.fairmlbook.org.

Beauchamp, T. L., Childress, J. F., et al. *Principles of Biomedical Ethics*. Oxford University Press, USA, 2001.

Brooks, J. P. Support vector machines with the ramp loss and the hard margin loss. *Operations Research*, 59(2): 467–479, 2011.

Chen, I., Johansson, F. D., and Sontag, D. Why Is My Classifier Discriminatory? In *Advances in Neural Information Processing Systems*, 2018.

Chen, I. Y., Szolovits, P., and Ghassemi, M. Can AI help reduce disparities in general medical and mental health care? *AMA Journal of Ethics*, 21(2):167–179, 2019.

Corbett-Davies, S. and Goel, S. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806. ACM, 2017.

Cotter, A., Gupta, M., Jiang, H., Srebro, N., Sridharan, K., Wang, S., Woodworth, B., and You, S. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. 2018.

Crawford, K. The hidden biases in big data. *Harvard Business Review*, 1, 2013.

Dastin, J. Amazon scraps secret ai recruiting tool that showed bias against women. *San Fransico, CA: Reuters. Retrieved on October*, 9:2018, 2018.

Dietterich, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

Dunn, O. J. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64, 1961.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. Decoupled classifiers for group-fair and efficient machine learning. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 119–133. PMLR, 2018.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268. ACM, 2015.

Ferryman, K. and Pitcan, M. Fairness in precision medicine, 2018.

Hardt, M., Price, E., Srebro, N., et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pp. 3315–3323, 2016.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. Fairness Without Demographics in Repeated Loss Minimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1929–1938. PMLR, 2018.

Hébert-Johnson, Ú., Kim, M., Reingold, O., and Rothblum, G. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the International Conference on Machine Learning*, pp. 1944–1953, 2018.

Hellman, D. *When is discrimination wrong?* Harvard University Press, 2008.

Hu, L. and Chen, Y. Fair Classification and Social Welfare. *arXiv preprint arXiv:1905.00147*, 2019.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2018.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Rambachan, A. Algorithmic Fairness. In *AEA Papers and Proceedings*, volume 108, pp. 22–27, 2018.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.

Lichman, M. UCI Machine Learning Repository, 2013.

Lipton, Z., McAuley, J., and Chouldechova, A. Does mitigating ml's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems 31*, pp. 8135–8145, 2018.

Nabi, R. and Shpitser, I. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1931, 2018.

Narasimhan, H. Learning with complex loss functions and constraints. In *International Conference on Artificial Intelligence and Statistics*, pp. 1646–1654, 2018.

National Lung Screening Trial Research Team. The National Lung Screening Trial: Overview and Study Design. *Radiology*, 258(1):243–253, 2011.

Nguyen, T. and Sanner, S. Algorithms for direct 0-1 loss optimization in binary classification. pp. 1085–1093, 2013.

Popejoy, A. B. and Fullerton, S. M. Genomics is failing on diversity. *Nature News*, 538(7624):161, 2016.

Salimi, B., Rodriguez, L., Howe, B., and Suciu, D. Capuchin: Causal database repair for algorithmic fairness. *arXiv preprint arXiv:1902.08283*, 2019.

Ustun, B. and Rudin, C. Supersparse Linear Integer Models for Optimized Medical Scoring Systems. *Machine Learning*, 102(3):349–391, 2016.

Ustun, B., Westover, M. B., Rudin, C., and Bianchi, M. T. Clinical prediction models for sleep apnea: the importance of medical history over symptoms. *Journal of Clinical Sleep Medicine*, 12(02):161–168, 2016.

Vayena, E., Blasimme, A., and Cohen, I. G. Machine learning in medicine: Addressing ethical challenges. *PLoS medicine*, 15(11):e1002689, 2018.

Wang, H., Ustun, B., and Calmon, F. P. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2019.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*, 2017.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1171–1180. International World Wide Web Conferences Steering Committee, 2017a.

Zafar, M. B., Valera, I., Rodriguez, M., Gummadi, K., and Weller, A. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pp. 228–238, 2017b.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 962–970. PMLR, 20–22 Apr 2017c.

Zeng, J., Ustun, B., and Rudin, C. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3):689–722, 2017.

Zhang, J. and Bareinboim, E. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems 31*, pp. 3675–3685. 2018.

Zink, A. and Rose, S. Fair Regression for Health Care Spending. *arXiv preprint arXiv:1901.10566*, 2019.

Zliobaite, I. A Survey on Measuring Indirect Discrimination in Machine Learning. *arXiv preprint arXiv:1511.00148*, 2015.

A. Proof of Theorem 2

In what follows, we present proofs of Theorem 2. We start a simple sufficient condition to ensure that a group prefers classifier h to another classifier h'. We will make use of this result to prove Theorem 2, and to design the score function for our decoupling procedure in Appendix B.

Lemma 3 (Generalization of Preferences) *Consider evaluating the true risk of two classifiers h and h' over group z. Given classifiers satisfy $\hat{\Delta}_z(h, h') > 0$, then $\Delta_z(h, h') > 0$ with probability at least $1 - \delta$ for any $\delta \in (0, 1]$ if*

$$4\mathfrak{R}(\mathcal{H}) + \sqrt{\frac{2\ln\frac{2}{\delta}}{n_z}} \leq \hat{\Delta}_z(h, h'), \tag{5}$$

where $\mathfrak{R}(\mathcal{H})$ is the Rademacher complexity of the hypothesis class $\mathcal{H}$.

Proof 1 *For any group $z \in Z$ and any classifier $h \in \mathcal{H}$ with probability at least $1 - \delta/2$, we have that*

$$\left|\hat{R}_z(h) - R_z(h)\right| \leq 2\mathfrak{R}(\mathcal{H}) + \sqrt{\frac{\ln\frac{2}{\delta}}{2n_z}}. \tag{6}$$

The bound in (6) holds for both h and h' with probability at least $1 - \delta$. Thus, we know that:

$$\begin{aligned}
R_z(h') - R_z(h) =& (R_z(h') - \hat{R}_z(h')) + (\hat{R}_z(h)) - R_z(h)) + \hat{R}_z(h') - \hat{R}_z(h) \\
\geq& -\left(2\mathfrak{R}(\mathcal{H}) + \sqrt{\frac{\ln\frac{2}{\delta}}{2n_z}}\right) - \left(2\mathfrak{R}(\mathcal{H}) + \sqrt{\frac{\ln\frac{2}{\delta}}{2n_z}}\right) + \hat{\Delta}_z(h, h') \\
=& -\left(4\mathfrak{R}(\mathcal{H}) + \sqrt{\frac{2\ln\frac{2}{\delta}}{n_z}}\right) + \hat{\Delta}_z(h, h') \\
\geq& 0,
\end{aligned}$$

if the condition specified in (5) holds.

We can make use of Lemma 3 to produce the following bounds on the generalization of rationality and envy-freeness. [6]

Corollary 4 (Generalization of Rationality) *Given a set of decoupled classifiers $H_Z = \{\hat{h}_z\}_{z \in Z}$ such that*

$$\hat{\Delta}_z(\hat{h}_z, \hat{h}_0) > 0 \quad \textit{for all} \quad z \in Z,$$

H_Z satisfies rationality with respect the pooled classifier $\hat{h}_0$ with probability at least $1 - \delta$, if for all groups $z \in Z$:

$$4\mathfrak{R}(\mathcal{H}) + \sqrt{\frac{2}{n_z}\ln\left(\frac{2|Z|}{\delta}\right)} \leq \hat{\Delta}_z(\hat{h}_z, \hat{h}_0),$$

Corollary 5 (Generalization of Envy-freeness) *Given a set of decoupled classifiers $H_Z = \{\hat{h}_z\}_{z \in Z}$ such that*

$$\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'}) > 0 \quad \textit{for all} \quad z, z' \in Z,$$

H_Z satisfies envy-freeness with probability at least $1 - \delta$ if, for all pairs of groups $z, z' \in Z$:

$$4\mathfrak{R}(\mathcal{H}) + \sqrt{\frac{2}{n_z}\ln\left(\frac{|Z|^2}{\delta}\right)} \leq \hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'}).$$

[6] For the sake of clarity, we will consider a setting where each group is assigned its own classifier so that $a(z) = z$ for each $z \neq z'$. Similar results can be derived for a setting where a single classifier can be assigned to multiple groups (see e.g., Appendix B).

Both results follow from repeated applications of Lemma 2. Specifically:

- Rationality requires that the pairwise preferences in Lemma 2 hold for all groups $z \in Z$. This involves preference conditions for $|Z|$ pairs of classifiers – i.e., one for each distinct pair $\hat{h}_z, \hat{h}_0$ where $z \in Z$. Thus, we can ensure that rationality holds with probability at least $1-\delta$ by applying Lemma 2 with probability at least $1 - \frac{\delta}{|Z|}$.
- Envy-freeness requires that the pairwise preferences in Lemma 2 hold for all pairs of groups $z, z' \in Z$. This involves preference conditions on $|Z|(|Z|-1)/2$ pairs of classifiers – i.e., one for each distinct pair $\hat{h}_z, \hat{h}_{z'}$ where $z, z' \in Z$. Since there are $|Z|(|Z|-1)/2$ pairs, and that $|Z|(|Z|-1)/2 \leq |Z|^2/2$, we can ensure that envy-freeness hold with probability at least $1-\delta$ by applying Lemma 2 with probability at least $\frac{\delta}{|Z|^2/2}$.

We are now ready to prove Theorem 2.

Proof 2 (Theorem 2) *Using Massart's Lemma, we have that:*

$$\mathfrak{R}(\mathcal{H}) \leq \sqrt{\frac{2\log|\mathcal{H}|}{n_z}} \tag{7}$$

Combining the bound on $\mathfrak{R}(\mathcal{H})$ in (7) with the bound in Corollary 4, we have that H_Z satisfies rationality with probability at least $1-\delta$, if for all $z \in Z$,

$$n_z \geq \frac{64\ln|\mathcal{H}| + 4\ln\left(\frac{2|Z|}{\delta}\right)}{\hat{\Delta}_z(\hat{h}_z, \hat{h}_0)^2} \tag{8}$$

Likewise, combining the bound on $\mathfrak{R}(\mathcal{H})$ in (7) with the bound in Corollary 5, we have that H_Z satisfies envy-freeness with probability at least $1-\delta$ if for all $z \in Z$,

$$n_z \geq \frac{64\ln|\mathcal{H}| + 4\ln\left(\frac{|Z|^2}{\delta}\right)}{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})^2}. \tag{9}$$

Given the bounds in (8) and (9), we can see that H_Z satisfies both rationality and envy-freeness with probability at least $1-\delta$ if for all $z \in Z$,

$$n_z \geq \max\left\{\frac{64\ln|\mathcal{H}| + 4\ln\left(\frac{2|Z|}{\delta}\right)}{\hat{\Delta}_z(\hat{h}_z, \hat{h}_0)^2}, \frac{64\ln|\mathcal{H}| + 4\ln\left(\frac{|Z|^2}{\delta}\right)}{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})^2}\right\} \tag{10}$$

Thus, the bound in Theorem 2 holds so long as we can show that:

$$\frac{64\ln|\mathcal{H}| + 4\ln(\frac{2|Z|^2}{\delta})}{\hat{\epsilon}_z^2} \geq \max\left\{\frac{64\ln|\mathcal{H}| + 4\ln\left(\frac{2|Z|}{\delta}\right)}{\hat{\Delta}_z(\hat{h}_z, \hat{h}_0)^2}, \frac{64\ln|\mathcal{H}| + 4\ln\left(\frac{|Z|^2}{\delta}\right)}{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})^2}\right\} \tag{11}$$

This follows given that we have defined $\hat{\epsilon}_z = \min\left(\hat{\Delta}_z(\hat{h}_z, \hat{h}_0), \min_{z' \in Z/\{z\}} \hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})\right)$, and that the inequality $4\ln\left(\frac{|Z|^2}{\delta}\right) \geq 4\ln\left(\frac{2|Z|}{\delta}\right)$ holds whenever $|Z| \geq 2$.

B. Score Function

In what follows, we formally derive the score function that we present in Section 4. The score function ensures that our procedure grows a tree in a way that is aligned with the goal of minimizing the risk of a preference violation.

We wish to bound the probability that H_T violates rationality or envy-freeness as follows:

$$\mathbb{P}\left(\begin{smallmatrix}H_T \text{ violates}\\ \text{rationality or}\\ \text{envy-freeness}\end{smallmatrix}\right) \leq \textsf{ViolationScore}(T) = \sum_{z \in Z} 4\exp\left(-\tfrac{n_z}{2} \cdot \hat{\Delta}_z(\hat{h}_z, \hat{h}_0)^2\right) + \sum_{z \in Z} \sum_{\substack{z' \in Z \\ a(z') \neq a(z)}} 4\exp\left(-\tfrac{n_z}{2} \cdot \hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})^2\right)$$

We restrict our attention to cases where $\hat{\Delta}_z(z, z') > 0$ since our training procedure ensures that $\hat{\Delta}_z(z, z') \geq 0$, and since $\hat{\Delta}_z(z, z') = 0$ implies indifference (i.e., it does not imply a preference violation).

Given a pair groups $z, z' \in Z$ such that $a(z) \neq a(z')$, we denote an event where group z prefers the classifier assigned to group z' as $\mathcal{E}_{z \to z'}$.. We will bound the probability of $\mathcal{E}_{z \to z'}$ in terms of the following event:

$$\mathcal{E}_{z,z'} = \left\{ |R_z(\hat{h}_z) - \hat{R}_z(\hat{h}_z)| \geq \frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})}{2} \right\} \cup \left\{ |R_z(\hat{h}_{z'}) - \hat{R}_z(\hat{h}_{z'})| \geq \frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})}{2} \right\}$$

We observe that $\mathcal{E}_{z \to z'} \subseteq \mathcal{E}_{z,z'}$. We proceed to present a proof by contradiction. Suppose that $\mathcal{E}_{z \to z'} \not\subseteq \mathcal{E}_{z,z'}$, this means that there must exist an event $\omega \in \mathcal{E}_{z \to z'}$ such that $\omega \notin \mathcal{E}_{z,z'}$. The fact that $\omega \notin \mathcal{E}_{z,z'}$ implies that both of the following inequalities must hold:

$$|R_z(\hat{h}_z) - \hat{R}_z(\hat{h}_z)| < \frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})}{2}$$
$$|R_z(\hat{h}_{z'}) - \hat{R}_z(\hat{h}_{z'})| < \frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})}{2}$$

This implies:

$$\begin{aligned} R_z(\hat{h}_z) - R_z(\hat{h}_{z'}) &= (R_z(\hat{h}_z) - \hat{R}_z(\hat{h}_z)) + (\hat{R}_z(\hat{h}_z) - \hat{R}_z(\hat{h}_{z'})) + (\hat{R}_z(\hat{h}_{z'}) - R_z(\hat{h}_{z'})) \\ &< \frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})}{2} - \hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'}) + \frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})}{2} \\ &= 0. \end{aligned}$$

Thus, we have shown that z does not envy z', which contradicts the fact that $\omega \in \mathcal{E}_{z \to z'}$.

Having shown that $\mathcal{E}_{z \to z'} \subseteq \mathcal{E}_{z,z'}$, we can bound the probability of an envy-freeness violation as follows:

$$\begin{aligned} \mathbb{P}\left(\cup_{z,z'} \mathcal{E}_{z \to z'}\right) &\leq \mathbb{P}\left(\cup_{z,z'} \mathcal{E}_{z,z'}\right) && (12) \\ &\leq \sum_{\substack{z,z' \in Z \\ a(z) \neq a(z')}} \mathbb{P}\left(\mathcal{E}_{z,z'}\right) && (13) \\ &\leq \sum_{\substack{z,z' \in Z \\ a(z) \neq a(z')}} \mathbb{P}\left(|R_z(\hat{h}_z) - \hat{R}_z(\hat{h}_z)| \geq \tfrac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})}{2}\right) + \mathbb{P}\left(|R_z(\hat{h}_{z'}) - \hat{R}_z(\hat{h}_{z'})| \geq \tfrac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})}{2}\right) && (14) \\ &\leq \sum_{\substack{z,z' \in Z \\ a(z) \neq a(z')}} 2\exp\left(-2n_z\left(\frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})}{2}\right)^2\right) + 2\exp\left(-2n_z\left(\frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})}{2}\right)^2\right) && (15) \\ &= \sum_{\substack{z,z' \in Z \\ a(z) \neq a(z')}} 4\exp\left(-\frac{n_z}{2} \cdot \hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})^2\right) && (16) \end{aligned}$$

Here: (12) follows from the fact that $\mathcal{E}_{z \to z'} \subseteq \mathcal{E}_{z,z'}$; (13) and (14) follow from the union bound; and (15) follows from inverting the bound.

We bound the probability of a rationality violation in a similar manner. We first define the following event for each $z \in Z$:

$$\mathcal{E}_{z,0} = \left\{ |R_z(\hat{h}_z) - \hat{R}_z(\hat{h}_z)| \geq \frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_0)}{2} \right\} \cup \left\{ |R_z(\hat{h}_0) - \hat{R}_z(\hat{h}_0)| \geq \frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_0)}{2} \right\}$$

We note that $\mathcal{E}_{z\to 0} \subseteq \mathcal{E}_{z,0}$, which can be shown by deriving an analogous contradiction to the one derived for envy-freeness. With this result, we can bound the probability of an rationality violation as follows:

$$\begin{aligned}
\mathbb{P}\left(\cup_{z\in Z}\mathcal{E}_{z\to 0}\right) &\leq \mathbb{P}\left(\cup_z \mathcal{E}_{z,0}\right) && (17)\\
&\leq \sum_{z\in Z} \mathbb{P}\left(\mathcal{E}_{z,0}\right) && (18)\\
&\leq \sum_{z\in Z} \mathbb{P}\left(\left(|R_z(\hat{h}_z) - \hat{R}_z(\hat{h}_z)| \geq \tfrac{\hat{\Delta}_z(\hat{h}_z,\hat{h}_0)}{2}\right) + \mathbb{P}\left(|R_z(\hat{h}_0) - \hat{R}_z(\hat{h}_0)| \geq \tfrac{\hat{\Delta}_z(\hat{h}_z,\hat{h}_0)}{2}\right)\right. && (19)\\
&\leq \sum_{z\in Z} 2\exp\left(-2n_z\left(\frac{\hat{\Delta}_z(\hat{h}_z,\hat{h}_0)}{2}\right)^2\right) + 2\exp\left(-2n_z\left(\frac{\hat{\Delta}_z(\hat{h}_z,\hat{h}_0)}{2}\right)^2\right) && (20)\\
&= \sum_{z\in Z} 4\exp\left(-\frac{n_z}{2}\cdot\hat{\Delta}_z(\hat{h}_z,\hat{h}_0)^2\right) && (21)
\end{aligned}$$

Here: (17) follows from the fact that $\mathcal{E}_{z\to 0} \subseteq \mathcal{E}_{z,0}$; (18) and (19) follow from the union bound; and (20) follows from inverting the bound. Our final expression for the score function is obtained by combining the terms in (16) and (21).