Minimax Pareto Fairness: A Multi Objective Perspective

Natalia Martinez [* 1] **Martin Bertran** [* 1] **Guillermo Sapiro** [1]

Abstract

In this work we formulate and formally characterize group fairness as a multi-objective optimization problem, where each sensitive group risk is a separate objective. We propose a fairness criterion where a classifier achieves minimax risk and is Pareto-efficient w.r.t. all groups, avoiding unnecessary harm, and can lead to the best zero-gap model if policy dictates so. We provide a simple optimization algorithm compatible with deep neural networks to satisfy these constraints. Since our method does not require test-time access to sensitive attributes, it can be applied to reduce worst-case classification errors between outcomes in unbalanced classification problems. We test the proposed methodology on real case-studies of predicting income, ICU patient mortality, skin lesions classification, and assessing credit risk, demonstrating how our framework compares favorably to other approaches.

1. Introduction

Machine learning algorithms play an important role in decision making in society. When these are used to make high-impact decisions such as hiring, credit-lending, predicting mortality for intensive care unit patients, or classifying skin lesions, it is paramount to guarantee that the prediction is both accurate and unbiased with respect to sensitive attributes such as gender or ethnicity. A model that is trained naively may not have these properties by default; see, for example (Barocas & Selbst, 2016).

In these critical applications, it is desirable to impose some fairness criteria. Some well-known definitions of group fairness in the machine learning literature attempt to make algorithms whose predictions are independent of the sensitive populations (e.g., Demographic Parity, (Louizos et al., 2015; Zemel et al., 2013; Feldman et al., 2015)); or al-

---
[*]Equal contribution [1]Department of Electircal and Computer Engineering, Duke University. Correspondence to: Natalia Martinez <natalia.martinez@duke.edu>.

gorithms whose outputs are independent of the sensitive attribute given the objective's ground truth (e.g., Equality of Odds, Equality of Opportunity, (Hardt et al., 2016; Woodworth et al., 2017)). Notions of Individual Fairness have also been proposed (Dwork et al., 2012; Joseph et al., 2016; Zemel et al., 2013). These can be appropriate in many scenarios, but in domains where quality of service is paramount, such as healthcare, we argue that it is necessary to strive for models that are as close to fair as possible without introducing unnecessary harm (Ustun et al., 2019). Additionally, a model satisfying these characteristics can be post-processed to introduce a controlled performance degradation that results in a perfectly fair, albeit harmful classifier. This is a decision beyond algorithmic design and is left to the policy-maker, but machine-learning should inform fairness policy and provide the necessary tools to implement it.

Here we focus on group fairness in terms of predictive risk disparities, a metric that has been explored in recent works such as (Calders & Verwer, 2010; Dwork et al., 2012; Feldman et al., 2015; Chen et al., 2018; Ustun et al., 2019). We formulate fairness as a Multi-Objective Optimization Problem and use Pareto optimality (Mas-Colell et al., 1995) to define the set of all efficient classifiers, meaning that the increase in predictive risk on one group is due to a decrease in the risk of another (no unnecessary harm). We consider problems where target labels available for training are trustworthy (not affected by discrimination), and tackle fairness as a minimax problem where the goal is to find the classifier with the smallest maximum group risk among all efficient models. As a design choice, this implies that a system's risk is as good as its worst group performance, but we do not enforce zero risk disparity if the disadvantaged groups do not benefit directly. When perfect fairness is achievable, this reduces to finding the efficient classifier in our hypothesis class that has the same risks among all groups. Our approach differs from post-hoc correction methods like the ones proposed in (Hardt et al., 2016; Woodworth et al., 2017), where zero-disparity is enforced by design, and test-time access to sensitive attributes is needed. Since our proposed methodology does not require the latter, and is not restricted to binary sensitive and target variables, it can also be used to reduce worst-case classification error between outcomes in imbalanced classification scenarios.

Main Contributions. We formulate group fairness as a Multi-Objective Optimization Problem (MOOP), where each objective function is the sensitive group conditional risk of the model. We formalize *no unnecessary harm* fairness using Pareto optimality (Mas-Colell et al., 1995); and characterize the space of Pareto-efficient classifiers for convex models and risk functions, which include deep neural networks (DNNs) and standard classifier losses. We show that all efficient classifiers under these conditions can be recovered with a simple modification of the overall risk function. We introduce and discuss minimax Pareto fairness (MMPF), where we select the efficient classifier with the smallest worst group conditional risk, and provide a simple and efficient algorithm to recover this classifier using standard (Stochastic) Gradient Descent on top of an adaptive loss. Critical to numerous applications in fairness and privacy, the proposed methodology does not require test-time access to the sensitive attributes. We also show that if the policy mandate is to obtain a zero-gap classifier, we can add harmful post-hoc corrections to the MMPF model, which ensures the lowest risk levels across all groups under certain conditions. In addition to this, we demonstrate how our methodology performs on real tasks such as inferring income status in the Adult dataset (Dua & Graff, 2017a), predicting ICU mortality rates in the MIMIC-III dataset from hospital notes (Johnson et al., 2016), classifying skin lesions in the HAM10000 dataset (Tschandl et al., 2018), and assessing credit risk on the German Credit dataset (Dua & Graff, 2017b). Finally, since our methodology does not require access test-time sensitive attributes, it can be used to reduce worst-case classification error between outcomes in unbalanced classification problems. Code is available at github.com/natalialmg/MMPF.

2. Related Work

There is a growing body of work on group fairness in machine learning. Following (Friedler et al., 2019), we empirically compare our methodology against the works of (Feldman et al., 2015; Kamishima et al., 2012; Zafar et al., 2015). Our method shares conceptual similarities with (Zafar et al., 2017; Woodworth et al., 2017; Agarwal et al., 2018; Oneto et al., 2019), but differs on the fairness objective and how it is adapted to work with standard neural networks. Although optimality is often discussed in the fairness literature, it is usually in the context of error-unfairness tradeoffs (Kearns & Roth, 2019; Kearns et al., 2017), and not between sensitive groups as studied here. The conflict between perfect fairness and optimality has been previously studied in (Kaplow & Shavell, 1999), we acknowledge this impossibility and formally characterize what is achievable in the context of machine learning and classification.

The work presented in (Hashimoto et al., 2018) discusses decoupled classifiers (one per sensitive group) as a way of minimizing group-risk disparity, but simultaneously cautions against this methodology when presented with insufficiently large datasets. The works of (Chen et al., 2018; Ustun et al., 2019) also empirically report the disadvantages of decoupled classifiers as a way to mitigate risk disparity. Here we argue for the use of a single classifier since it does not require access to sensitive group membership during test time, and might allow transfer learning between diverse groups when possible. If access to group membership during test time is available, this can be naturally incorporated as part of our observation features; with a sufficiently rich hypothesis set, this is equivalent to training separate classifiers, with the added benefit of positive transfer on samples were groups share optimal decision boundaries (Ustun et al., 2019; Wang et al., 2020).

The work of (Chen et al., 2018) uses the unified bias-variance decomposition advanced in (Domingos, 2000) to identify that noise levels across different sub-populations may differ, making perfect fairness parity impossible without explicitly degrading performance on one group. Their methodology attempts to bridge the disparity gap by collecting additional samples from high-risk sub-populations. Here we modify the classifier loss to improve worst-case group performance without inducing unnecessary harm, which could be considered synergistic with their methodology.

3. Minimax Pareto Fairness: Formulation and Basic Properties

Consider we have access to a dataset $\mathcal{D} = \{(x_i, y_i, a_i)\}_{i=1}^n$ containing n independent triplet samples drawn from a joint distribution $(x_i, y_i, a_i) \sim P(X, Y, A)$, where $x_i \in \mathcal{X}$ are our input features (e.g., images, tabular data), $y_i \in \mathcal{Y}$ is our target variable, and $a_i \in \mathcal{A}$ indicates group membership or sensitive status (e.g., ethnicity, gender); our input features X may or may not explicitly contain A, meaning sensitive attributes need not be available at deployment.

Let $h \in \mathcal{H}$ be a classifier from a hypothesis class $\mathcal{H}$ trained to infer y from x, $h : \mathcal{X} \to [0, 1]^{|\mathcal{Y}|}$. We use $\delta^Y \in \{0, 1\}^{|\mathcal{Y}|} : \delta_i^Y = \mathbb{1}(Y = y_i), i = 1, ..., |\mathcal{Y}|$, to denote the one-hot representation of Y. Given a loss function $\ell : [0, 1]^{|\mathcal{Y}|} \times [0, 1]^{|\mathcal{Y}|} \to \mathbb{R}^+$ the group-specific risk of classifier h on group a is $r_a(h) = E_{X,Y|A=a}[\ell(h(X), \delta^Y)]$. We approach fairness as a Multi-Objective Optimization Problem (MOOP), where the classifier h is our decision variable, the group-specific risks $\{r_a(h)\}_{a=1}^{|\mathcal{A}|}$ are our objective functions, and they conform a risk vector $\boldsymbol{r}(h) = \{r_a(h)\}_{a=1}^{|\mathcal{A}|}$. The MOOP can be stated as

$$\min_{h \in \mathcal{H}} (r_1(h), r_2(h), \ldots, r_{|\mathcal{A}|}(h)). \tag{1}$$

We use dominance (Miettinen, 2008) to define optimality for a MOOP, namely, Pareto optimality, the definitions are given below. We later formally characterize the space of

optimal multi-objective classifiers h for well-known losses, and argue a fairness criteria where a fair classifier is both Pareto optimal and has the smallest maximum group risk. Lemmas and theorems are stated without proof throughout the main text, proofs are provided in Section A.1.

Definition 3.1. Dominant vector: A vector $\boldsymbol{r}' \in \mathbb{R}^k$ is said to dominate $\boldsymbol{r} \in \mathbb{R}^k$, noted as $\boldsymbol{r}' \prec \boldsymbol{r}$, if $r_i' \leq r_i, \forall i = 1, \ldots, k$ and $\exists j : r_j' < r_j$ (i.e., strict inequality on at least one component). Likewise, we denote $\boldsymbol{r}' \preceq \boldsymbol{r}$ if $\boldsymbol{r} \not\prec \boldsymbol{r}'$

Definition 3.2. Dominant classifier: Given a set of group-specific risk functions $\boldsymbol{r}(h)$, a classifier h' is said to dominate h'', noted as $h' \prec h''$, if $\boldsymbol{r}(h') \prec \boldsymbol{r}(h'')$. Similarly, we denote $h' \preceq h''$ if $\boldsymbol{r}(h') \preceq \boldsymbol{r}(h'')$.

Definition 3.3. Pareto front and Pareto optimality: Given a family of classifiers $\mathcal{H}$, and a set of group-specific risk functions $\boldsymbol{r}(h)$, the set of Pareto front classifiers is $\mathcal{P}_{\mathcal{A},\mathcal{H}} = \{h \in \mathcal{H} : \nexists h' \in \mathcal{H}|h' \prec h\} = \{h \in \mathcal{H} : h \preceq h' \forall h' \in \mathcal{H}\}$. The corresponding achievable risks are denoted as $\mathcal{P}^{\mathcal{R}}_{\mathcal{A},\mathcal{H}} = \{\boldsymbol{r} \in \mathbb{R}^{+|\mathcal{A}|} : \exists h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}, \boldsymbol{r} = \boldsymbol{r}(h)\}$. A classifier h is a Pareto optimal solution to the MOOP in Eq.(1) iff $h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}$.

No unnecessary harm fairness. The Pareto front defines the best achievable trade-offs between population risks $r_a(h)$. This is already suited for classification and regression tasks where the sensitive attributes are categorical. Constraining the classifier to be in the Pareto front disallows laziness, there exists no other classifier in the hypothesis class $\mathcal{H}$ that is at least as good on all group-specific risks and strictly better in one of them. In this sense, we say that a classifier in the Pareto front does *no unnecessary harm*.

Literature on fairness has focused on putting constraints on the norm of discrimination gaps (Zafar et al., 2017; 2015; Creager et al., 2019; Woodworth et al., 2017). Here we focus on minimizing the risk on the worst performing group (Definition 3.4), these two criteria often yield similar results, and can be shown to be identical for Pareto optimal classifiers when $|\mathcal{A}| = 2$. For more than 2 sensitive groups, there may be situations where minimum risk discrepancy leads to higher minimax risk (e.g., a classifier that increases both the minimum and maximum risk but decreases their gap is still optimal if a third group sees their risk diminished). Constraining solutions to be Pareto optimal and minimizing the maximum risk preserves the overall idea of reducing risk disparities while avoiding some potentially undesirable tradeoffs. We formalize this next:

Definition 3.4. Minimax Pareto fair classifier and Minimax Pareto fair vector: A classifier h^* is a Minimax Pareto fair classifier if it minimizes the worst group-specific risk among all Pareto front classifiers, $h^* \in \underset{h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}}{\arg\min} \underset{a \in \mathcal{A}}{\max} r_a(h) = \underset{h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}}{\arg\min} \|\boldsymbol{r}(h)\|_\infty$, with corresponding Minimax Pareto-fair risk vector $\boldsymbol{r}^* = \boldsymbol{r}(h^*)$.

An important consequence of this formulation is that when the hypothesis class $\mathcal{H}$ contains a classifier that is both Pareto optimal and has zero risk disparity, this classifier is also Minimax Pareto fair.

Lemma 3.1. *If* $\exists h^* \in \mathcal{P}_{\mathcal{A},\mathcal{H}} : r_a(h^*) = r_{a'}(h^*), \forall a, a' \in \mathcal{A}$ *then* $\boldsymbol{r}(h^*) = \arg\min_{\boldsymbol{r} \in \mathcal{P}^{\mathcal{R}}_{\mathcal{A},\mathcal{H}}} \|\boldsymbol{r}\|_\infty$.

Even when perfect equality of risk is desirable, Pareto classifiers still serve as useful intermediaries. To this end, Lemma 3.2, shows that any classifier in $\mathcal{H}$ that attains equality of risk has worse performance on all groups than the Minimax Pareto fair classifier. Furthermore, we can post-process the Pareto fair classifier to be perfectly fair by increasing risk on over-performing groups, this procedure is still no worse than obtaining an equal risk classifier in our hypothesis class.

Lemma 3.2. *Let* $h_{ER} \in \mathcal{H}$ *be an equal risk classifier such that* $r_a(h_{ER}) = r_{a'}(h_{ER}) \forall a, a'$, *and let* h^* *be the Pareto fair classifier. Additionally, define the Pareto fair post-processed equal risk classifier* $h^*_{ER} : r_a(h^*_{ER}) = \|\boldsymbol{r}(h^*)\|_\infty \forall a \in \mathcal{A}$, *then we have*

$$r_a(h_{ER}) \geq r_a(h^*_{ER}) \geq r_a(h^*) \ \forall a \in \mathcal{A}.$$

We provide a fair optimal classifier that improves the worst group risk. It also serves as an intermediate step to get perfect fairness; the decision between the two is left to the policymaker. To exemplify these notions graphically, Figure 1 shows a scenario with binary sensitive attributes a where none of the Pareto classifiers achieve equality of risk. Here the noise level differs between groups, and the Pareto fair risk $\boldsymbol{r}^*$ is not achieved by either a Naive classifier (minimizes expected global risk), or a classifier where groups are re-sampled to appear with equal probability (Balanced classifier). We observe how the discrimination gap along the Pareto front is closed by trading off performance from one group to another. The gap can be further closed by moving outside the Pareto front, but this discrimination reduction is a result of performance degradation on the privileged group, with no tangible upside to the underprivileged one.

Section 5 provides a method to recover the minimax Pareto fair classifier (MMPF) from training samples. Before that, Section 4 shows important properties of Pareto-efficient classifiers for convex hypothesis classes, including DNNs, and risk functions.

4. Analysis of Pareto Optimal Solutions

In this section we characterize the Pareto front for convex hypothesis classes and convex risk functions. under these conditions the models have attractive regularity properties.

Definition 4.1. Convex hypothesis class and risk function: A hypothesis class $\mathcal{H} \subseteq \{h : \mathcal{X} \rightarrow [0,1]^{|\mathcal{Y}|}\}$ is convex iff $\forall h', h'' \in \mathcal{H}, \ \lambda \in [0,1], \rightarrow \lambda h' + (1-\lambda)h'' \in \mathcal{H}$.

A risk function $r : \mathcal{H} \rightarrow \mathbb{R}_+$ is convex iff $\forall \lambda \in [0,1] \ r(\lambda h' + (1-\lambda)h'') \leq \lambda r(h') + (1-\lambda)r(h'')$
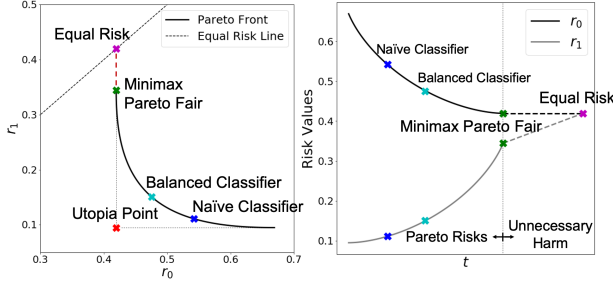
Figure 1. Achievable risk trade-offs for a binary classification problem $Y \in \{0, 1\}$ with two unbalanced groups $A \in \{0, 1\}$, and covariates $X|A \sim N(\mu_A, 1)$ (parameters provided in Supplementary Material). *Left: Pareto front risks and the equal risk line; the minimax Pareto fair point (green) does not achieve equality of risk; the trade-offs attained by standard (Naive, blue) classifier and a class-rebalanced (Balanced, cyan) classifier are also shown. The Utopia point (red) corresponds to the minimum achievable risk for each group. Right: Parametrization of group risks along the Pareto front line, and on the minimax Pareto fair to Equal Risk line (Unnecessary harm). All points in the Pareto front efficiently trade-off performance between groups; the trajectory outside of the Pareto front, however, does not improve performance on the worst performing group r_0, it only degrades performance on r_1.*

Definition 4.2. Convex Pareto front: A Pareto front $\mathcal{P}^{\mathcal{R}}_{\mathcal{A},\mathcal{H}} \subseteq \mathbb{R}^{|\mathcal{A}|}$ is convex if $\forall r, r' \in \mathcal{P}^{\mathcal{R}}_{\mathcal{A},\mathcal{H}}, \lambda \in [0, 1], \exists r^\lambda \in \mathcal{P}^{\mathcal{R}}_{\mathcal{A},\mathcal{H}}$ such that $r^\lambda \preceq \lambda r + (1-\lambda) r'$

Convexity of the hypothesis class for DNNs can be seen as a natural consequence of the Universal Function Approximation Theorem, shown for fully connected neural networks in (Hornik et al., 1989), and more recently for convolutional NNs (CNNs) in (Zhou, 2020). Note that this convexity is w.r.t. its function output space (i.e., for any two classifiers in the hypothesis class h_1, h_2, there exists a classifier in this same family $h_\lambda : h_\lambda(x) = (1 - \lambda)h_1(x) + \lambda h_2(x) \forall x, \lambda \in [0, 1]$), the parameter space itself may be highly non-convex. As for risk functions, many standard classification losses such as Brier Score ($r^{BS}_a(h) = E_{X,Y|a}[||\delta^Y - h(X)||^2_2]$) and Cross Entropy ($r^{CE}_a(h) = E_{X,Y|a}[\langle \delta^Y, \ln(h(X)) \rangle]$) are convex w.r.t. the classifier output.

The following theorem (Theorem 4.1) shows that under these conditions, the Pareto front can be fully characterized by solving the linear weighting problem:

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \sum_{a=1}^{|\mathcal{A}|} \mu_a r_a(h);$$
$$\|\boldsymbol{\mu}\|^1_1 = 1, \ \mu_a > 0, \ a = 1, ..., |\mathcal{A}|. \quad (2)$$

We use the shorthand notation $h^{\boldsymbol{\mu}}$ to describe a classifier that solves Problem 2; likewise, we denote $r(\boldsymbol{\mu}) = r_{h^{\boldsymbol{\mu}}}$. We utilize the results derived in (Geoffrion, 1968) to show that when both the hypothesis class $\mathcal{H}$ and the risk functions are convex, any optimal classifier $h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}$ is a solution to Problem 2 for some choice of weights $\boldsymbol{\mu}$. Note that even

when the risk functions and hypothesis class are non-convex, solutions to Problem 2 still belong to the Pareto front.

Theorem 4.1. Given $\mathcal{H}$ a convex hypothesis class and $\{r_a(h)\}_{a \in \mathcal{A}}$ convex risk functions then:

1. The Pareto front is convex: $\forall r, r' \in \mathcal{P}^{\mathcal{R}}_{\mathcal{A},\mathcal{H}}, \lambda \in [0, 1], \exists r'' \in \mathcal{P}^{\mathcal{R}}_{\mathcal{A},\mathcal{H}} : r'' \preceq \lambda r + (1 - \lambda) r'$.

2. Every Pareto solution is a solution to Problem 2: $\forall \hat{r} \in \mathcal{P}^{\mathcal{R}}_{\mathcal{A},\mathcal{H}}, \exists \boldsymbol{\mu} : \hat{r} = r(\boldsymbol{\mu})$.

We can then characterize the Pareto optimal classifiers for Brier score (BS) and Cross-Entropy (CE) in the infinite samples and unbounded hypothesis class regime. The following provides an expression for the optimal classifiers and risks in terms of the probability densities and weights.

Theorem 4.2. Given input features $X \in \mathcal{X}$, categorical target $Y \in \mathcal{Y}$ and sensitive group $A \in \mathcal{A}$, with joint distribution $p(X, Y, A)$, and weights $\boldsymbol{\mu} = \{\mu_a\}_{a \in \mathcal{A}}$, the optimal predictor to the linear weighting problem $h(\boldsymbol{\mu})$ for both Brier score and Cross-Entropy is

$$h^{\boldsymbol{\mu}}(x) = \frac{\sum_{a \in \mathcal{A}} \mu_a p(x|a) p(y|x, a)}{\sum_{a \in \mathcal{A}} \mu_a p(x|a)},$$

with corresponding risks

$$r^{BS}_a(\boldsymbol{\mu}) = E_{X,Y|a}[||\delta^Y - p(y|X, a)||^2_2] +$$
$$E_{X|a}\left[||p(y|X, a) - h^{\boldsymbol{\mu}}(X)||^2_2\right],$$
$$r^{CE}_a(\boldsymbol{\mu}) = H(Y|X, a) +$$
$$E_{X|a}\left[D_{KL}\left(p(y|X, a) || h^{\boldsymbol{\mu}}(X)\right)\right],$$

where $p(y|X, a) = \{p(Y = y_i | X, A = a)\}^{|\mathcal{Y}|}_{i=1}$ is the probability mass vector of Y given X and $A = a$. $H(Y|X, a)$ is the conditional entropy $H(Y|X, A = a) = E_{X|A=a}[H(Y|X = X, A = a)]$.

The optimal risks for BS and CE are decomposed as the sum of two non-negative terms. The first term corresponds to the minimum achievable group risk, attained with the group-specific optimal classifier $p(y|X, a)$, this is independent of $h^{\boldsymbol{\mu}}$. The second term measures the discrepancy between $p(y|X, a)$ and the optimal predictor $h^{\boldsymbol{\mu}}$. Since both risk functions are convex, this is a full characterization of all asymptotically optimal multi-objective classifiers; this includes our proposed minimax Pareto fair classifier.

From the expressions in Theorem 4.2 we observe that if the separability condition $Y \perp A|X$ is satisfied, the minimum risk for each subgroup is attained. Here the Pareto front only contains the Utopia point (see Figure 1). Additionally, if the entropy of the sensitive attribute given our features is small (A is well predicted from X), the Pareto front also tends to the Utopia point. This is formalized in the following lemma.

Lemma 4.3. *In the conditions of Theorem 4.2 we observe that if $Y \perp A|X$ then*

$$r^{BS}_a(\boldsymbol{\mu}) = E_{X,Y|a}[||\delta^Y - p(y|X)||^2_2] \ \forall \boldsymbol{\mu},$$
$$r^{CE}_a(\boldsymbol{\mu}) = H(Y|X) \ \forall \boldsymbol{\mu}.$$

Likewise, if $H(A|X) \to 0$ then
$$r_a^{BS}(\boldsymbol{\mu}) \to E_{X,Y|a}[||\delta^Y - p(y|X,a)||_2^2] \; \forall \boldsymbol{\mu},$$
$$r_a^{CE}(\boldsymbol{\mu}) \to H(Y|X,a) \; \forall \boldsymbol{\mu}.$$

Note that even on these ideal cases ($Y \perp A \mid X$ or $H(A|X) \to 0$), the baseline risks between groups might differ. Therefore, perfect equality of risk may only be achieved by selecting sub-optimal classifiers (unnecessary harm), or by improving the input features X; this observation concurs with previous analysis on classifiers' bias-variance tradeoffs done in (Chen et al., 2018; Domingos, 2000).

5. Minimax Pareto Fair Optimization

Our goal is to find the Pareto classifier h^* that minimizes the risk of the worst performing sensitive groups (i.e., $h^* \in \arg\min_{h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}} ||\boldsymbol{r}(h)||_\infty$). It is important to note that the classifier h^* is not necessarily unique, nor is its corresponding risk vector $\boldsymbol{r}^* = \boldsymbol{r}(h^*)$. Throughout this section, we assume that the hypothesis for Theorem 4.1 are satisfied, therefore, for every minimax vector $\boldsymbol{r}^*$ there is a set of weights $\boldsymbol{\mu}^*$ such that $\boldsymbol{r}^*$ is a unique solution for Problem 2 ($\boldsymbol{r}^* = \boldsymbol{r}(\boldsymbol{\mu}^*)$).

Computing $\boldsymbol{\mu}^*$ directly can be challenging, even when closed form solutions for the classifiers and risks are available, as shown in Theorem 4.2 for Brier Score and Cross Entropy. A potential approach to estimate $\boldsymbol{\mu}^*$ would be to perform sub-gradient descent on $||\boldsymbol{r}(\boldsymbol{\mu})||_\infty$. This approach suffers from two main setbacks. First, closed form formulas for the Jacobian $\nabla_{\boldsymbol{\mu}} \boldsymbol{r}(\boldsymbol{\mu})$ require accurate estimates of the conditional distributions $p(x|a)$ and $p(y|x,a)$; or of $p(a), p(a|x)$ and $p(y|x,a)$. Secondly, $||\boldsymbol{r}(\boldsymbol{\mu})||_\infty$ can potentially have local minima on $\boldsymbol{\mu}$.

We propose a simple optimization method to recover $\boldsymbol{\mu}^*$ that only requires access to function evaluations of $\boldsymbol{r}(\boldsymbol{\mu})$. Note that given $\boldsymbol{\mu}$, the risk vector $\boldsymbol{r}(\boldsymbol{\mu})$ can be obtained by estimating the optimal classifier with the expression derived in Theorem 4.2 (plug-in estimation), which requires the conditional densities $p(a), p(a|x)$ and $p(y|x,a)$. Another option is to directly minimize Problem, 2 (joint estimation). Both approaches can be implemented using DNNs for estimation. Note that the second approach makes use of all samples to estimate a single classifier, while estimating the densities $p(a|x), p(y|x,a)$ individually may suffer from data fragmentation and cannot benefit from transfer learning, which could be harmful for a sensitive group with limited data (Ustun et al., 2019; Wang et al., 2020).

To avoid blind sampling of the weighting vectors $\boldsymbol{\mu}$, Theorem 5.1 summarizes important properties that any weighting vector $\boldsymbol{\mu}'$ must satisfy to improve the minimax risk at any given iteration ($\boldsymbol{\mu}' : ||\boldsymbol{r}(\boldsymbol{\mu}')||_\infty < ||\boldsymbol{r}(\boldsymbol{\mu})||_\infty$).

Theorem 5.1. Let $\mathcal{P}_{\mathcal{A},\mathcal{H}}^{\mathcal{R}}$ be a Pareto front, and $\boldsymbol{r}(\boldsymbol{\mu}) \in \mathcal{P}_{\mathcal{A},\mathcal{H}}^{\mathcal{R}}$ denote the solution to the linear weighting Problem 2. For any $\boldsymbol{\mu}' \notin \arg\min_{\boldsymbol{\mu} \in \Delta^{|\mathcal{A}|-1}} ||\boldsymbol{r}(\boldsymbol{\mu})||_\infty$, and $\boldsymbol{\mu}^* \in$

$\arg\min_{\boldsymbol{\mu} \in \Delta^{|\mathcal{A}|-1}} ||\boldsymbol{r}(\boldsymbol{\mu})||_\infty$, the sets $N_i = \{\boldsymbol{\mu} : r_i(\boldsymbol{\mu}) < ||\boldsymbol{r}(\boldsymbol{\mu}')||_\infty\}$ satisfy:

1. $\boldsymbol{\mu}^* \in \bigcap_{i \in \mathcal{A}} N_i$;

2. If $\boldsymbol{\mu} \in N_i \to \lambda \boldsymbol{\mu} + (1-\lambda)\boldsymbol{e}^i \in N_i, \forall \lambda \in [0,1], i = 1, \ldots, |\mathcal{A}|$, where $\boldsymbol{e}^i$ denotes the standard basis vector;

3. $\forall \mathcal{I} \subseteq \mathcal{A}, \boldsymbol{\mu} : \mu_{\mathcal{A} \setminus \mathcal{I}} = 0 \to \boldsymbol{\mu} \in \bigcup_{i \in \mathcal{I}} N_i$;

4. If $\boldsymbol{r}(\boldsymbol{\mu})$ is also continuous in $\boldsymbol{\mu}$, then $\forall \mathcal{I} \subseteq \mathcal{A}$ such that $\boldsymbol{\mu} \in \bigcap_{i \in \mathcal{I}} N_i \to \exists \epsilon > 0 : B_\epsilon(\boldsymbol{\mu}) \subset \bigcap_{i \in \mathcal{I}} N_i$;

5. If $\mathcal{P}_{\mathcal{A},\mathcal{H}}^{\mathcal{R}}$ is also convex, then $\boldsymbol{r}(\boldsymbol{\mu}^*) \in \arg\min_{\boldsymbol{r} \in \mathcal{P}_{\mathcal{A},\mathcal{H}}^{\mathcal{R}}} ||\boldsymbol{r}||_\infty$.

Therefore, finding an updated weighting vector $\boldsymbol{\mu}$ that diminishes the minimax risk is equivalent to finding an element in the intersection of $|\mathcal{A}|$ subsets defined on the $\Delta^{|\mathcal{A}|-1}$ simplex. These subsets N_i are themselves star-shaped sets w.r.t. the basis element $\boldsymbol{e}^i$, whose intersections $\cap_{i \in \mathcal{I}} N_i$ form open sets. Property 3 in the theorem provides a straightforward rule to find elements belonging to any arbitrary union of the coordinate descent regions $\cup_{i \in \mathcal{I}} N_i$.
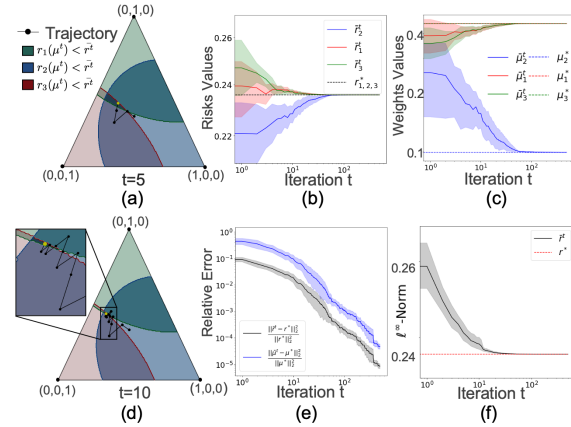
Figure 2. Synthetic data experiment with 3 sensitive groups. (a) and (d) show a simplex diagram of the linear weights $\boldsymbol{\mu}$ on the fifth and tenth iteration of the APStar algorithm; blue, green and red shaded areas correspond to the N_i areas at iterations, $t = 5, t = 10$, the optimal linear weight lies in their intersection. (b) and (c) show the risk values and linear weights as a function of the iteration counter; shaded regions represent standard deviations across 5 randomized runs. (e) shows relative error as a function of iterations for both risks and weights; (f) shows similar information comparing the maximum risk against the theoretical optimal. Risks for all groups converge to the minimax value, while the weights converge to $\boldsymbol{\mu}^*$. Simulation details are provided in Section A.4.

Using these observations, we propose the Approximate Projection onto Star Sets (APStar) Algorithm (Algorithm 1) to iteratively refine the minimax risk by updating the linear weighting vector. The main intuition behind this algorithm is that whenever we observe a weighting vector that

Algorithm 1 APStar

Input: hypothesis class: $\mathcal{H}$, initial weights: $\boldsymbol{\mu}$, risk functions: $r_a(\cdot)$, optimizer: $\{\arg\} \min_{h \in \mathcal{H}} \sum_{i=1}^{|\mathcal{A}|} \mu_i r_i(h)$, $\alpha \in (0,1)$, K_{min}

Initialize:
$h, \boldsymbol{r}(\mu) \leftarrow \{\arg\} \min_{h \in \mathcal{H}} \sum \mu_i r_i(h)$
$\bar{\boldsymbol{r}} \leftarrow \|\boldsymbol{r}(\mu)\|_\infty$; $K \leftarrow 1$.
repeat
$\quad \mathbf{1}_\mu \leftarrow \{\mathbf{1}(r_i(\mu) \geq \bar{r})\}_{i=1}^{|\mathcal{A}|}$
$\quad \boldsymbol{\mu} \leftarrow (\alpha\boldsymbol{\mu} + \frac{1-\alpha}{K\|\mathbf{1}_\mu\|_1^1}\mathbf{1}_\mu)\frac{K}{(K-1)\alpha+1}$
$\quad h, \boldsymbol{r}(\mu) \leftarrow \{\arg\} \min_{h \in \mathcal{H}} \sum \mu_i r_i(h)$; $K \leftarrow K+1$
$\quad$ **if** $\|\boldsymbol{r}(\mu)\|_\infty < \bar{r}$ **then**
$\quad\quad \bar{\boldsymbol{r}} \leftarrow \|\boldsymbol{r}(\mu)\|_\infty$, $K \leftarrow \min(K, K_{min})$
$\quad\quad h^*, \boldsymbol{\mu}^*, \boldsymbol{r}^* \leftarrow h, \boldsymbol{\mu}, \boldsymbol{r}(\mu)$
$\quad$ **end if**
until $Convergence$
Return: $h^*, \boldsymbol{\mu}^*, \boldsymbol{r}^*$

reduces the risk in a subset of groups $\mathcal{I}$ ($\boldsymbol{\mu}^t \in \cap_{i \in \mathcal{I}} N_i$), we can generate a new vector $\boldsymbol{\mu}^{t+1}$ that linearly interpolates between $\boldsymbol{\mu}^t$ and a vector $\boldsymbol{\mu}^{\mathcal{A} \setminus \mathcal{I}} \in \cup_{i \in \mathcal{A} \setminus \mathcal{I}} N_i$ that belongs to the union of the unsatisfied group risks ($\boldsymbol{\mu}^{t+1} \to \alpha\boldsymbol{\mu}^t + (1-\alpha)\boldsymbol{\mu}^{\mathcal{A} \setminus \mathcal{I}}$). An analysis of the convergence properties of APStar is provided in Section A.2.

Figure 2 illustrates how the linear weights $\boldsymbol{\mu}$ are updated on a synthetic example, reducing the minimax risk; the example shown is of a classifier with 3 sensitive groups where perfect fairness is attainable, we observe how risks converge to their common final value $\boldsymbol{r}^*$ and that the weights $\boldsymbol{\mu}^*$ required to recover this are not equally-weighted vector.

In Section 6 we apply the APStar algorithm to synthetic and real datasets using DNNs and SGD to solve Problem 2. Note that APStar can be applied to non-convex risk functions and hypothesis classes, in which case the solutions to Problem 2 form a subset of the Pareto front and may not offer a full characterization of all optimal solutions. In Section A.3 we provide implementation details; Pytorch code for this algorithm will be made available.

6. Experiments and Results

<mark>We applied the proposed APStar algorithm to learn a minimax Pareto fair classifier (MMPF) and show how our approach produces well calibrated models that improve minimax performance across several metrics beyond the risk measure itself.</mark> While details are provided in Section A.2, Figure 3 illustrates how our algorithm is empirically convergent and significantly faster than random sampling and the multiplicative weight update (MWU) algorithm proposed in (Chen et al., 2017) for minimax optimization in the context of robustness. In sections A.4 and A.5 we validate the optimization algorithm on synthetic data and compare joint versus plug-in estimation. Here we compare the performance

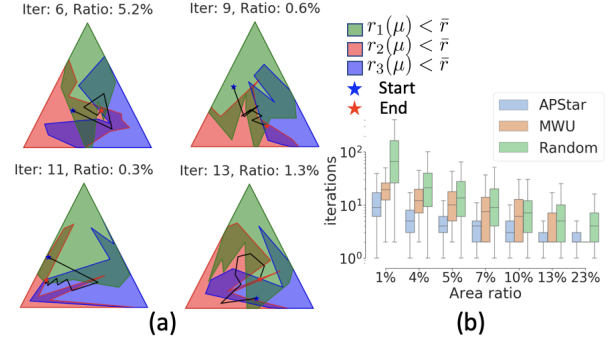of our method against other state of the art approaches on a variety of public fairness datasets.

Figure 3. Synthetic data experiment on star-shaped sets. (a) Randomly sampled star sets satisfying the conditions of Theorem 5.1; a starting point is sampled (Blue), trajectories recovered by the APStar algorithm are recorded until convergence (Red); number of iterations and intersection area are shown for all examples. (b) Empirical distribution of number of iterations required to converge versus percentage of linear weights that lie in the triple intersection; values are shown for APStar, random sampling, and the multiplicative weight update (MWU) algorithm proposed in (Chen et al., 2017) for minimax optimization. The number of iterations required by the algorithm is well below the random sampler, this is especially apparent for low area ratio scenarios.

6.1. Real Datasets: Methods and Metrics

We evaluate our method on mortality prediction (MIMIC-III), skin lesion classification (HAM10000), income prediction (Adult), and credit lending (German). The latter two are common benchmarks in the fairness literature. Results are reported for joint estimation (MMPF) and plug-in estimation (MMPF P), presented in Section 5. We evaluate a model trained to minimize the average risk (Naive) and one that samples all <mark>sensitive groups equally (Balanced).</mark> When applicable, we compare our results against the methodologies proposed in (Hardt et al., 2016; Zafar et al., 2015; Kamishima et al., 2012; Feldman et al., 2015); for implementations on all methods except (Hardt et al., 2016), we used the unified test-bed provided in (Friedler et al., 2019). A description of these methods is provided in Section A.6.

Metrics used for evaluation include accuracy (Acc), Brier Score (BS) and Cross-Entropy (CE), values are reported per sensitive attribute, standard deviations computed across 5 splits are shown when available. For datasets containing more than two sensitive groups (MIMIC-III, HAM10000) we report dataset average (sample mean), group-normalized average (group mean), worst performing group (worst group), <mark>and largest difference between groups (disparity). Note that the latter two are especially important in our setting, since our explicit focus is to minimize worst-case performance,</mark> and disparity is a common measure of interest in fairness literature. For an in-depth description of these metrics and datasets, refer to Sections A.7 and A.8. Additional

tables showing these and other metrics on a per-group basis are provided in Section A.10. All tables bold the best result for disparity and worst group risk.

Similarly to (Kamishima et al., 2012; Zafar et al., 2015; Hardt et al., 2016; Woodworth et al., 2017), we omit the sensitive attribute from our observation features, which broadens the potential application of the framework. Classifiers are implemented using neural networks and/or linear logistic regression; for details on architectures and hyperparameters, refer to Section A.9.

6.2. Predicting Mortality in Intensive Care Patients

We used clinical notes collected from adult ICU patients at the Beth Israel Deaconess Medical Center (MIMIC-III dataset) (Johnson et al., 2016) to predict patient mortality. We study fairness with respect to age (adult/senior), ethnicity (white/nonwhite), and outcome (alive/deceased) simultaneously (8 sensitive groups). Outcome is included as a sensitive attribute because, in our experiments, patients who ultimately passed away on ICU were under-served by a Naive classifier (high classifier loss). Using the target label as sensitive attribute addresses class imbalance by requiring similar risk performance on each target label. It also demonstrates a use-case where group membership would not be available at test-time.

We use a fully connected NN with BS loss as the base hypothesis class (results with CE provided in Section A.10), input features are Tf-idf statistics on the $10k$ most frequent words from clinical notes, mirroring (Chen et al., 2018). For an even comparison, we provided the feature embeddings of the Naive classifier as input to the baselines, since the implementation in (Friedler et al., 2019) only includes linear classifiers, this was also done since some of the available implementations failed to converge in a reasonable time with the original inputs. Table 1 reports Acc and BS of tested methodologies. Note that, while the Balanced classifier has significantly better worst-case BS performance than the Naive classifier, MMPF is better still; these performance gains are also reflected on Acc, showing that this improvement goes beyond the training metric. Plug-in performance (MMPF P) is not an improvement over joint estimation. Accuracy comparison table incorporates the post-processing methodology described in (Hardt et al., 2016) to achieve zero accuracy disparity ("+H" suffix). Hardt post-processing decreases the accuracy disparity gap w.r.t. the baseline methods but requires test-time access to group membership. The best result is obtained on MMPF H, but we note that the best minimax risk is still attained on the original MMPF model.

6.3. Skin Lesion Classification

The HAM10000 dataset (Tschandl et al., 2018) contains over $10k$ dermatoscopic images of 7 types of skin lesions; with ratios between 67% and 1.1%. A Naive classifier exhib-

Table 1. MIMIC dataset. Group priors range from 0.4% to 57%. In this and all tables we bold the best result for disparity and worst group risk.

(a) Acc comparison

	Sample mean	Group mean	Worst group	Disparity
Naive	89.5 ±0.2	61.9 ±1.7	19.0 ±2.0	80.5 ±1.3
Balanced	79.4±0.6	77.5±1.4	66.8±2.2	22.6±2.3
Zafar	86.2±0.3	65.8±1.8	32.0±2.4	62.9±3.6
Feldman	88.6±2.4	64.4±2.9	28.7±2.4	72.1±5.5
Kamishima	89.3±0.2	63.6±2.0	25.1±5.1	76.4±5.2
MMPF	76.2±0.2	78.3±1.5	**72.6±1.7**	**17.1±3.5**
MMPF P	75.5±1.0	76.8±1.3	70.7±2.1	17.8±3.8
Balanced+H	75.6±1.1	71.7±1.6	65.6±2.8	19.1±1.8
Zafar+H	62.8±1.6	58.3±2.1	51.5±2.8	17.8±3.1
MMPF+H	72.4±1.1	72.3±1.5	**72.0±3.7**	**11.4±3.5**

(b) BS comparison

	Sample mean	Group mean	Worst group	Disparity
Naive	.16 ±.01	.51 ±.02	1.05 ±.01	1.03 ±.02
Balanced	.28±.01	.31±.01	.42±.01	.25±.04
Zafar	.27±.01	.67±.04	1.34±.05	1.25±.07
Feldman	.19±.04	.62±.04	1.26±.08	1.29±.08
Kamishima	.16±.01	.53±.03	1.06±.04	1.11±.08
MMPF	.32±.01	.3±.01	**.35±.02**	**.17±.03**
MMPF P	.33±.01	.32±.01	.37±.01	**.17±.04**

ited no significant discrimination based on age or race. We instead chose to use the diagnosis class as both the target and sensitive variable. It was not possible to compare against (Hardt et al., 2016) since the sensitive attribute is perfectly predictive of the outcome; likewise, (Zafar et al., 2015) and (Kamishima et al., 2012) cannot handle non-binary target attributes in their provided implementations. Table 2 shows results for MMFP P, since plug-in estimation is equivalent to joint estimation when $A = Y$, but enables cheaper APStar iterations (see Section A.5). The MMPF classifier improves minimax BS and Acc when compared to both Naive and Balanced classifiers.

Table 2. HAM10000 dataset. Group priors range from 1% to 67%.

(a) Acc comparison

	Sample mean	Group mean	Worst group	Disparity
Naive	78.5 ± 0.5	50.8 ± 1.9	2.6 ±3.5	93.7 ±1.1
Balanced	70.1 ±2.1	70.1 ±2.2	52.6 ±5.3	32.5 ± 4.6
MMPF P	64.7 ±1.2	66.7 ± 3.5	**56.9 ±3.1**	**19.8 ±6.6**

(b) BS comparison

	Sample mean	Group mean	Worst group	Disparity
Naive	.31±.01	.69±.3	1.38±.04	1.29±.04
Balanced	.41±.02	.42±.03	0.64±.05	0.45±.07
MMPF P	.49±.02	.46±.04	**0.56±.4**	**0.23±.06**

6.4. Income Prediction and Credit Risk Assesment

We predict income in the Adult UCI dataset (Dua & Graff, 2017a) and assess credit risk in the German Credit dataset (Dua & Graff, 2017b). We select gender (Male/Female) as our sensitive attribute, additional results for gender and ethnicity (Male/Female and White/Other) are also shown for the Adult dataset. Tables 3, 4, and 5 show results for linear logistic regression (LR suffix) and a fully connected NN. Note that linear logistic regression is not a convex hypothesis class, but is included to compare evenly against the baselines.

Our approach leads to the best worst-case performance in Acc and CE on the Adult dataset, although all methods perform similarly; this is especially true for the gender case, where Kamishima has a slight advantage in terms of standard deviation in CE. On the German dataset, our method produces the best worst-case CE results, and smallest disparities in both Acc and CE; Feldman does, however, have test time access to sensitive attributes, which may explain the difference in Acc. We show results for Hardt post-processing in Section A.10, with similar conclusions to the ones made on the MIMIC dataset.

Table 3. Adult gender dataset. Females represent 32% of samples.

(a) Acc comparison

	Female	Male	Disparity
Naive LR	92.3±0.4	80.5±0.4	11.9±0.7
Balanced LR	92.3±0.3	80.3±0.7	12.0±0.7
Zafar	92.5±0.3	80.9±0.3	11.6±0.4
Feldman	92.3±0.3	80.7±0.2	11.6±0.1
Kamishima	92.6±0.4	80.9±0.4	11.7±0.7
MMPF LR	91.9±0.4	81.0±0.4	10.9±0.7
MMPF	92.1±0.3	81.3±0.3	10.8±0.5
MMPF LR P	92.0±0.4	81.0±0.5	11.0±0.6
MMPF P	91.7±0.3	**81.5±0.5**	**10.1±0.5**

(b) CE comparison

	Female	Male	Disparity
Naive LR	.204±.009	.411±.006	.207±.007
Balanced LR	.204±.011	.416±.011	.211±.005
Zafar	.202±.018	.398±.006	.195±.023
Feldman	.201±.004	.403±.004	.203±.006
Kamishima	.189±.006	**.395±.004**	.206±.007
MMPF LR	.204±.008	**.395±.006**	.19±.011
MMPF	.21±.019	.403±.025	.193±.013
MMPF LR P	.208±.008	**.395±.005**	.187±.01
MMPF P	.227±.019	.403±.023	**.176±.014**

Table 4. Adult ethnicity and gender dataset. Group priors range from 6% to 60%.

(a) Acc comparison

	Sample mean	Group mean	Worst group	Disparity
Naive LR	84.7±0.3	87.8±0.1	80.6±0.5	14.1±1.0
Balanced LR	84.7±0.3	88.0±0.3	80.5±0.5	14.5±1.0
Zafar	84.7±0.2	87.9±0.2	80.6±0.5	14.5±0.9
Feldman	84.5±0.2	87.7±0.3	80.4±0.3	14.7±0.9
Kamishima	84.3±0.8	87.8±0.3	80.0±1.2	15.2±1.8
MMPF LR	84.5±0.2	87.8±0.3	80.6±0.5	14.0±1.0
MMPF	84.8±0.3	87.8±0.4	80.9±0.6	13.6±1.5
MMPF LR P	84.6±0.3	87.7±0.2	80.7±0.5	13.9±1.0
MMPF P	84.6±0.4	87.6±0.5	**81.0±0.8**	**13.4±1.5**

(b) CE comparison

	Sample mean	Group mean	Worst group	Disparity
Naive LR	.332±.004	.268±.004	.408±.008	.268±.016
Balanced LR	.333±.004	.268±.005	.411±.008	.273±.015
Zafar	.334±.005	.273±.005	.409±.01	.266±.03
Feldman	.337±.003	.276±.006	.412±.006	.262±.016
Kamishima	.337±.015	.275±.009	.414±.023	.269±.026
MMPF LR	.334±.004	.274±.005	**.404±.007**	**.251±.015**
MMPF	.334±.005	.272±.003	.404±.009	.263±.022
MMPF LR P	.335±.005	.275±.006	.405±.006	**.251±.01**
MMPF P	.345±.009	.284±.01	.41±.014	.258±.03

Table 5. German dataset. Females represent 30% of samples.

(a) Acc comparison

	Female	Male	Disparity
Naive LR	70.7±7.3	71.2±4.5	8.8±4.7
Balanced LR	71.6±5.9	70.9±4.1	5.8±3.6
Zafar	73.0±5.6	71.0±3.5	5.8±3.5
Feldman	73.5±8.6	**71.9±4.3**	7.9±4.4
Kamishima	68.8±6.8	72.7±2.6	6.0±4.4
MMPF LR	72.5±5.5	71.6±2.8	5.0±2.6
MMPF LR P	70.7±4.5	71.5±3.6	**4.4±0.5**

(b) CE comparison

	Female	Male	Disparity
Naive LR	.607±.1	.559±.069	.127±.064
Balanced LR	.594±.082	.568±.068	.096±.05
Zafar	.567±.09	.735±.205	.273±.151
Feldman	.564±.096	.551±.063	.091±.068
Kamishima	.62±.064	.545±.062	.075±.067
MMPF LR	.565±.04	.544±.046	**.048±.041**
MMPF LR P	**.563±.043**	.537±.051	.057±.034

7. Discussion

Here we formulate group fairness as a multi objective optimization problem where each group-specific risk is an objective function. Our goal is to recover an efficient classifier that reduces worst-case group risks ethically (i.e., avoiding unnecessary harm). We consider problems where target labels available for training are trustworthy (not affected by discrimination). We formally characterized Pareto optimal solutions for a family of models and risk functions, yielding insight on the fundamental sources of risk trade-offs. We proposed a simple algorithm to recover a model that improves minimax group risk (MMPF), and does not require test-time access to sensitive attributes.

We demonstrated the proposed framework and optimization algorithm on several real-world case studies, achieving state-of-the-art performance. The algorithm is straightforward to implement, and is agnostic to the hypothesis class, risk function and optimization method, which allows integration with a variety of classification pipelines, including

neural networks. If the hypothesis class or risk functions are not convex, the algorithm can still be deployed to recover Pareto-efficient models that reduce minimax risks, though minimax optimality might not be achievable by optimizing a linear weighting of the risk functions. While this paper addresses minimal harm, other considerations like marginal risk tradeoffs between groups may be of interest. Controlling these in the proposed framework can be achieved by adding a constraint on the ratio between linear weights.

As an avenue of future research, we would like to automatically identify high-risk sub-populations as part of the learning process and attack risk disparities as they arise, rather than relying on preexisting notions of disadvantaged groups. The APStar algorithm is empirically convergent, but a formal proof or counterexample is desireable. We strongly believe that Pareto-efficient notions of fairness are of great interest for several applications, especially so on domains such as healthcare, where quality of service is paramount.

Acknowledgments

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.

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

Calders, T. and Verwer, S. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

Chen, I., Johansson, F. D., and Sontag, D. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems*, pp. 3539–3550, 2018.

Chen, R. S., Lucier, B., Singer, Y., and Syrgkanis, V. Robust optimization for non-convex objectives. In *Advances in Neural Information Processing Systems*, pp. 4705–4714, 2017.

Creager, E., Madras, D., Jacobsen, J.-H., Weis, M. A., Swersky, K., Pitassi, T., and Zemel, R. Flexibly fair representation learning by disentanglement. *arXiv preprint arXiv:1906.02589*, 2019.

Domingos, P. A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning*, pp. 231–238, 2000.

Dua, D. and Graff, C. UCI machine learning repository, 2017a. URL http://archive.ics.uci.edu/ml.

Dua, D. and Graff, C. UCI machine learning repository, 2017b. URL https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data).

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

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.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 329–338. ACM, 2019.

Geoffrion, A. M. Proper efficiency and the theory of vector maximization. *Journal of mathematical analysis and applications*, 22(3):618–630, 1968.

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. B., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. *arXiv preprint arXiv:1806.08010*, 2018.

Hornik, K., Stinchcombe, M., White, H., et al. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

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

Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. Rawlsian fairness for machine learning. *arXiv preprint arXiv:1610.09559*, 1(2), 2016.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50. Springer, 2012.

Kaplow, L. and Shavell, S. The conflict between notions of fairness and the pareto principle. *American Law and Economics Review*, 1(1):63–77, 1999.

Kearns, M. and Roth, A. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, 2019.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

Mas-Colell, A., Whinston, M. D., Green, J. R., et al. *Microeconomic theory*, volume 1. Oxford university press New York, 1995.

Miettinen, K. Introduction to multiobjective optimization: Noninteractive approaches. In *Multiobjective optimization*, pp. 1–26. Springer, 2008.

Oneto, L., Doninini, M., Elders, A., and Pontil, M. Taking advantage of multitask learning for fair classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 227–237, 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Tschandl, P., Rosendahl, C., and Kittler, H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5:180161, 2018.

Ustun, B., Liu, Y., and Parkes, D. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pp. 6373–6382, 2019.

Wang, H., Hsu, H., Diaz, M., and Calmon, F. P. To split or not to split: The impact of disparate treatment in classification. *arXiv preprint arXiv:2002.04788*, 2020.

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., Rodriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.

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. 229–239, 2017.

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

Zhou, D.-X. Universality of deep convolutional neural networks. *Applied and computational harmonic analysis*, 48 (2):787–794, 2020.

A. Supplementary Material

A.1. Proofs

Here we provide proof outlines for all lemmas and theorems in sections 3,4, and 5.

Lemma 3.1. If $\exists h^* \in \mathcal{P}_{\mathcal{A},\mathcal{H}} : r_a(h^*) = r_{a'}(h^*), \forall a, a' \in \mathcal{A}$ then $\boldsymbol{r}(h^*) = \underset{\boldsymbol{r} \in \mathcal{P}^{\mathcal{R}}_{\mathcal{A},\mathcal{H}}}{\arg\min} \|\boldsymbol{r}\|_\infty$.

Proof. By contradiction, assume $\exists h' \in \mathcal{P}_{\mathcal{A},\mathcal{H}} : \|\boldsymbol{r}(h')\|_\infty < \|\boldsymbol{r}(h^*)\|_\infty$. Then we have

$$r_a(h') \leq \|\boldsymbol{r}(h')\|_\infty < \|\boldsymbol{r}(h^*)\|_\infty = r_a(h^*), \ \forall a \in \mathcal{A}$$

And therefore $\boldsymbol{r}_{h'} \prec \boldsymbol{r}_{h*} \rightarrow h* \notin \mathcal{P}_{\mathcal{A},\mathcal{H}}$, which contradicts the hypothesis. $\square$

Lemma 3.2. Let $h_{ER} \in \mathcal{H}$ be an equal risk classifier such that $r_a(h_{ER}) = r_{a'}(h_{ER}) \forall a, a'$, and let h^* be the Pareto fair classifier. Additionally, define the Pareto fair post-processed equal risk classifier $h^*_{ER} : r_a(h^*_{ER}) = \|\boldsymbol{r}(h^*)\|_\infty \forall a \in \mathcal{A}$, then we have

Proof. Note that $\|\boldsymbol{r}_{h_{ER}}\|_\infty \geq \|\boldsymbol{r}_{h^*}\|_\infty$, otherwise $\|\boldsymbol{r}_{h_{ER}}\|_\infty < \|\boldsymbol{r}_{h^*}\|_\infty$ and also $h_{ER} \preceq h^*$, which contradicts the definition of h^*.

The statement therefore follows from $r_a(h_{ER}) = \|\boldsymbol{r}_{h_{ER}}\|_\infty \geq \|\boldsymbol{r}_{h^*}\|_\infty = r_a(h^*_{ER}) \geq r_a(h^*) \ \forall a \in \mathcal{A}$. $\square$

Theorem 4.1. Given $\mathcal{H}$ a convex hypothesis class and $\{r_a(h)\}_{a \in \mathcal{A}}$ convex risk functions then:

1. The Pareto front is convex: $\forall \boldsymbol{r}, \boldsymbol{r}' \in \mathcal{P}^{\mathcal{R}}_{\mathcal{A},\mathcal{H}}, \lambda \in [0,1], \exists \boldsymbol{r}'' \in \mathcal{P}^{\mathcal{R}}_{\mathcal{A},\mathcal{H}} : \boldsymbol{r}'' \preceq \lambda \boldsymbol{r} + (1-\lambda)\boldsymbol{r}'$.

2. Every Pareto solution is a solution to Problem 2: $\forall \hat{\boldsymbol{r}} \in \mathcal{P}^{\mathcal{R}}_{\mathcal{A},\mathcal{H}}, \exists \boldsymbol{\mu} : \hat{\boldsymbol{r}} = \boldsymbol{r}(\boldsymbol{\mu})$.

Proof. We prove the first item of the theorem statement, the second item is a direct application of the results in (Geoffrion, 1968).

Let $h', h'' \in \mathcal{P}_{\mathcal{A},\mathcal{H}}$, with corresponding risk vectors $\boldsymbol{r}_{h'}, \boldsymbol{r}_{h''}$. Using the convexity of $r_a(h), \forall \lambda \in [0,1]$ we have

$$\lambda r_a(h') + (1-\lambda)r_a(h') \geq r_a(\lambda h + (1-\lambda)h'').$$

Since $\mathcal{H}$ is convex, $h^\lambda = \lambda h' + (1-\lambda)h'' \in \mathcal{H}$. We have two possibilities;

- $h^\lambda \in \mathcal{P}_{\mathcal{A},\mathcal{H}}$, therefore, by definition $\boldsymbol{r}_{h^\lambda} \in \mathcal{P}^{\mathcal{R}}_{\mathcal{A},\mathcal{H}}$ and $\boldsymbol{r}_{h^\lambda} \preceq \lambda \boldsymbol{r}' + (1-\lambda)\boldsymbol{r}''$;;

- $h^\lambda \notin \mathcal{P}_{\mathcal{A},\mathcal{H}}$, therefore $\exists \hat{\boldsymbol{r}} \in \mathcal{P}^{\mathcal{R}}_{\mathcal{A},\mathcal{H}}, \hat{\boldsymbol{r}} \prec \boldsymbol{r}^\lambda$.

In both cases, for all risk vectors $\boldsymbol{r}', \boldsymbol{r}'' \in \mathcal{P}^{\mathcal{R}}_{\mathcal{A},\mathcal{H}}, \lambda \in [0,1] \exists \hat{\boldsymbol{r}} \in \mathcal{P}^{\mathcal{R}}_{\mathcal{A},\mathcal{H}} : \hat{\boldsymbol{r}} \preceq \lambda \boldsymbol{r}' + (1-\lambda)\boldsymbol{r}''$

$\square$

Theorem 4.2. Given input features $X \in \mathcal{X}$ and categorical target and sensitive group variables $Y \in \mathcal{Y}$ and $A \in \mathcal{A}$ respectively, with joint distribution $p(X, Y, A)$, and linear weights $\boldsymbol{\mu} = \{\mu_a\}_{a \in \mathcal{A}}$, the optimal predictor to the linear weighting problem $h(\boldsymbol{\mu})$ for both Brier score and Cross-Entropy is

$$h^{\boldsymbol{\mu}}(x) = \frac{\sum_{a \in \mathcal{A}} \mu_a p(x|a)p(y|x,a)}{\sum_{a \in \mathcal{A}} \mu_a p(x|a)},$$

with corresponding risks

$$\begin{aligned} r_a^{BS}(\boldsymbol{\mu}) &= E_{X,Y|a}[\|\delta^Y - p(y|X,a)\|_2^2] + \\ &\quad E_{X|a}\left[\|p(y|X,a) - h^{\boldsymbol{\mu}}(X)\|_2^2\right], \\ r_a^{CE}(\boldsymbol{\mu}) &= H(Y|X,a) + \\ &\quad E_{X|a}\left[D_{KL}\left(p(y|X,a)\|h^{\boldsymbol{\mu}}(X)\right)\right], \end{aligned}$$

where $p(y|X,a) = \{p(Y = y_i|X, A = a)\}_{i=1}^{|\mathcal{Y}|}$ is the probability mass vector of Y given X and $A = a$. $H(Y|X,a)$ is the conditional entropy $H(Y|X, A = a) = E_{X|A=a}[H(Y|X = X, A = a)]$.

Proof. We observe that the linear loss function $\sum_{a \in \mathcal{A}} \mu_a r_a(h)$ can be decomposed as

$$\begin{aligned} &\sum_{a \in \mathcal{A}} \mu_a r_a(h) \\ &= \sum_{a \in \mathcal{A}} \mu_a E_{X|a}\left[E_{Y|X,a}[\ell(h(X), Y)]\right] \\ &= \sum_{a \in \mathcal{A}} \mu_a E_X\left[\frac{p(X|a)}{p(X)}E_{Y|X,a}[\ell(h(X), Y)]\right] \\ &= \sum_{a \in \mathcal{A}} \mu_a E_X\left[\frac{p(X|a)}{p(X)}E_{Y|X}\left[\frac{p(Y|X,a)}{p(Y|X)}\ell(h(X), Y)\right]\right] \\ &= E_X\left[\frac{1}{p(X)}E_{Y|X}\left[\frac{\sum_{a \in \mathcal{A}} \mu_a p(X|a)p(Y|X,a)}{p(Y|X)}\ell(h(X), Y)\right]\right] \\ &= E_X\left[\frac{\sum_{a \in \mathcal{A}} \mu_a p(X|a)}{p(X)}E_{Y \sim P^\mu(Y|X)}[\ell(h(X), Y)]\right], \end{aligned}$$

with $P^\mu(Y|X) = \frac{\sum_{a \in \mathcal{A}} \mu_a p(X|a)p(y|X,a)}{\sum_{a \in \mathcal{A}} \mu_a p(X|a)}$, and denoting $p(y|X,a) = \{p(Y = y_i|X,a)\}_{i=1}^{|\mathcal{Y}|}$ the conditional probability mass vector of $Y|X, a$.

For both the Cross-Entropy loss $\ell^{CE}(h(X), Y) = \langle \delta^Y, \ln(h(X)) \rangle$ and the Brier score loss $\ell^{BS}(h(X), Y) = \|\delta^Y - h(X)\|_2^2$, the minimizer of $E_{Y \sim P(Y|X)}[\ell(h(X), Y)]$ is attained at $h(X) = P(Y|X)$. Therefore, we can plug in this optimal estimator to recover

$$h^{\boldsymbol{\mu}}(x) = \frac{\sum_{a \in \mathcal{A}} \mu_a p(x|a) p(y|x,a)}{\sum_{a \in \mathcal{A}} \mu_a p(x|a)}.$$

Plugging in the optimal classifier $h^{\boldsymbol{\mu}}(x)$ on the risk formulations we get the expressions for both scores as analyzed next.

Brier Score:

$$
\begin{aligned}
r_a^{BS}(\boldsymbol{\mu}) &= \\
&= E_{X,Y|a}[||\delta^Y - h^{\boldsymbol{\mu}}(X)||_2^2] \\
&= E_{X,Y|a}[||\delta^Y - p(y|X,a) + p(y|X,a) - h^{\boldsymbol{\mu}}(X)||_2^2] \\
&= E_{X,Y|a}[||\delta^Y - p(y|X,a)||_2^2] \\
&\quad + E_{X,Y|a}[||p(y|X,a) - h^{\boldsymbol{\mu}}(X)||_2^2] + \\
&\quad + 2E_{X,Y|a}\Big[[\delta^Y - p(y|X,a)]^T [p(y|X,a) - h^{\boldsymbol{\mu}}(X)]\Big] \\
&= r_a^{BSmin} + E_{X|a}\Big[||p(y|X,a) - h^{\boldsymbol{\mu}}(X)||_2^2\Big] \\
&\quad + 2E_{X|a}\Big[[p(y|X,a) - p(y|X,a)]^T\Big][p(y|X,a) - h^{\boldsymbol{\mu}}(X)] \\
&= r_a^{BSmin} + E_{X|a}\Big[||p(y|X,a) - h^{\boldsymbol{\mu}}(X)||_2^2\Big],
\end{aligned}
$$

where the equalities are attained by expanding the l2-norm, observing that $p(y|X,a)$ and $h(X)$ do not depend on random variable Y, and also that $E_{Y|X,a}[\delta^Y] = p(y|X,a)$. We also denoted $r_a^{BSmin} = E_{X,Y|a}[||\delta^Y - p(y|X,a)||_2^2]$, which is the risk attained by the Bayes optimal estimator for group a.

Cross Entropy:

$$
\begin{aligned}
r_a^{CE}(\boldsymbol{\mu}) &= \\
&= -E_{X,Y|a}\Big[\sum_{i=1}^{|\mathcal{Y}|} \delta_i^Y log[h_i^{\boldsymbol{\mu}}(X)]\Big] \\
&= -E_{X,Y|a}\Big[\sum_{i=1}^{|\mathcal{Y}|} \delta_i^Y log[\frac{h_i^{\boldsymbol{\mu}}(X) p(y_i|X,a)}{p(y_i|X,a)}]\Big] \\
&= -E_{X,Y|a}\Big[\sum_{i=1}^{|\mathcal{Y}|} \delta_i^Y log[p(y_i|X,a)]\Big] + \\
&\quad + E_{X,Y|a}\Big[\sum_{i=1}^{|\mathcal{Y}|} \delta_i^Y log[\frac{p(y_i|X,a)}{h_i^{\boldsymbol{\mu}}(X)}]\Big] = \\
&= r_a^{CEmin} + E_{X|a}\Big[\sum_{i=1}^{|\mathcal{Y}|} p(y_i|X,a) log[\frac{p(y_i|X,a)}{h_i^{\boldsymbol{\mu}}(X)}]\Big] \\
&= r_a^{CEmin} + E_{X|a}\Big[D_{KL}\Big(p(y|X,a)||h^{\boldsymbol{\mu}}(X)\Big)\Big],
\end{aligned}
$$

where we again use the linearity of the expectation and the equality $E_{Y|X,a}[\delta^Y] = p(y|X,a)$. We also denote $r_a^{CEmin} = H(Y|X, A = a)$. $\qquad\square$

Lemma 4.3. In the conditions of Theorem 4.2 we observe that if $Y \perp A|X$ then

$$
\begin{aligned}
r_a^{BS}(\boldsymbol{\mu}) &= E_{X,Y|a}[||\delta^Y - p(y|X)||_2^2] \;\forall \boldsymbol{\mu} \\
r_a^{CE}(\boldsymbol{\mu}) &= H(Y|X) \;\forall \boldsymbol{\mu},
\end{aligned}
$$

Likewise, if $H(A|X) \to 0$ then

$$
\begin{aligned}
r_a^{BS}(\boldsymbol{\mu}) &\to E_{X,Y|a}[||\delta^Y - p(y|X,a)||_2^2] \;\forall \boldsymbol{\mu} \\
r_a^{CE}(\boldsymbol{\mu}) &\to H(Y|X,a) \;\forall \boldsymbol{\mu}.
\end{aligned}
$$

Proof.
If $Y \perp A|X$ then

$$
\begin{aligned}
h^{\boldsymbol{\mu}}(x) &= \frac{\sum_{a \in \mathcal{A}} \mu_a p(x|a) p(y|x,a)}{\sum_{a \in \mathcal{A}} \mu_a p(x|a)}, \\
&= p(y|x),
\end{aligned}
$$

in which case the resulting expressions for $r_a^{BS}(\boldsymbol{\mu}), r_a^{CE}(\boldsymbol{\mu})$ are immediate and do not depend on $\boldsymbol{\mu}$

If $H(A|X) \to 0$ then we have $p(a|x) = 1, p(a'|x) = 0 \,\forall\, a \neq a', x : p(x|a) > 0$. Therefore we can write

$$
\begin{aligned}
h^{\boldsymbol{\mu}}(x) &= \frac{\sum_{a \in \mathcal{A}} \mu_a p(x|a) p(y|x,a)}{\sum_{a \in \mathcal{A}} \mu_a p(x|a)} \\
&= \frac{\sum_{a \in \mathcal{A}} \mu_a p(a|x) p(a) p(y|x,a)}{\sum_{a \in \mathcal{A}} \mu_a p(a|x) p(a)} \\
&= p(y|x,a), \;\forall a, x : p(x|a) > 0,
\end{aligned}
$$

and again the resulting expressions for $r_a^{BS}(\boldsymbol{\mu}), r_a^{CE}(\boldsymbol{\mu})$ are immediate from direct substitution. $\qquad\square$

We now present two auxiliary lemmas that will help us prove Theorem 5.1.

Lemma A.1. *Let $\mathcal{P}_{\mathcal{A},\mathcal{H}}^{\mathcal{R}}$ be a Pareto front, and let $r(\boldsymbol{\mu}) \in \mathcal{P}_{\mathcal{A},\mathcal{H}}^{\mathcal{R}}$ denote the solution the linear weighting Problem given by Eq.(2). Then $\forall \boldsymbol{\mu}, \in \mathbb{R}_+^{\mathcal{A}}, \mathcal{I} \subseteq \mathcal{A}, \boldsymbol{\eta} : \eta_i > 0 \forall i \in \mathcal{I}, \eta_i = 0 \forall i \in \mathcal{A} \setminus \mathcal{I}.$*

Then at least one risk in the $\mathcal{I}$ coordinates is reduced, or both risk vectors are the same in $\mathcal{I}$, i.e.,

$$\exists j \in \mathcal{I} : r_j(\boldsymbol{\mu} + \boldsymbol{\eta}) < r_j(\boldsymbol{\mu}) \vee \boldsymbol{r}_{\mathcal{I}}(\boldsymbol{\mu} + \boldsymbol{\eta}) = \boldsymbol{r}_{\mathcal{I}}(\boldsymbol{\mu}).$$

Proof. Denote $\Phi(\boldsymbol{\mu}) = \sum_{a \in \mathcal{A}} \mu_a r_a(\boldsymbol{\mu}), \; \Phi_{\mathcal{A}\setminus\mathcal{I}}(\boldsymbol{\mu}) = \sum_{a \in \mathcal{A}\setminus\mathcal{I}} \mu_a r_a(\boldsymbol{\mu})$, and $\boldsymbol{r}_{\mathcal{I}}(\boldsymbol{\mu}) = \{r_a(\boldsymbol{\mu})\}_{a \in \mathcal{I}}$.

By contradiction, we negate the thesis to get

$$(\boldsymbol{r}_{\mathcal{I}}(\boldsymbol{\mu} + \boldsymbol{\eta}) \geq \boldsymbol{r}_{\mathcal{I}}(\boldsymbol{\mu})) \wedge (\exists j \in \mathcal{I} : \boldsymbol{r}_j(\boldsymbol{\mu} + \boldsymbol{\eta}) > \boldsymbol{r}_j(\boldsymbol{\mu})).$$

We discuss 2 cases.

Case $\Phi_{\mathcal{A}\setminus\mathcal{I}}(\boldsymbol{\mu}) > \Phi_{\mathcal{A}\setminus\mathcal{I}}(\boldsymbol{\mu} + \boldsymbol{\eta})$:

By definition of $\boldsymbol{r}(\boldsymbol{\mu}), \boldsymbol{r}(\boldsymbol{\mu} + \boldsymbol{\eta})$ we observe

$$
\begin{aligned}
&\sum_{a \in \mathcal{A}}(\mu_a + \eta_a) r_a(\boldsymbol{\mu}) \geq \sum_{a \in \mathcal{A}}(\mu_a + \eta_a) r_a(\boldsymbol{\mu} + \boldsymbol{\eta}), \\
&\sum_{a \in \mathcal{I}}(\mu_a + \eta_a) r_a(\boldsymbol{\mu}) + \Phi_{\mathcal{A}\setminus\mathcal{I}}(\boldsymbol{\mu}) \geq \\
&\qquad \geq \sum_{a \in \mathcal{I}}(\mu_a + \eta_a) r_a(\boldsymbol{\mu} + \boldsymbol{\eta}) + \Phi_{\mathcal{A}\setminus\mathcal{I}}(\boldsymbol{\mu} + \boldsymbol{\eta}), \\
&\underbrace{\sum_{a \in \mathcal{I}}(\mu_a + \eta_a)(r_a(\boldsymbol{\mu}) - r_a(\boldsymbol{\mu} + \boldsymbol{\eta}))}_{\leq 0} \geq \\
&\qquad \geq \underbrace{\Phi_{\mathcal{A}\setminus\mathcal{I}}(\boldsymbol{\mu} + \boldsymbol{\eta}) - \Phi_{\mathcal{A}\setminus\mathcal{I}}(\boldsymbol{\mu})}_{>0},
\end{aligned}
$$

where the inequalities in the underbrackets are a direct result of the case assumptions, these inequalities contradict the hypothesis.

Case $\Phi_{\mathcal{A}\setminus\mathcal{I}}(\boldsymbol{\mu}) \leq \Phi_{\mathcal{A}\setminus\mathcal{I}}(\boldsymbol{\mu} + \boldsymbol{\eta})$:

We can directly observe that

$$
\begin{aligned}
\sum_{a\in\mathcal{A}}(\mu_a + \eta_a)r_a(\boldsymbol{\mu}) \\
= \sum_{a\in\mathcal{I}}(\mu_a + \eta_a)r_a(\boldsymbol{\mu}) + \Phi_{\mathcal{A}\setminus\mathcal{I}}(\boldsymbol{\mu}) \\
\leq \sum_{a\in\mathcal{I}}(\mu_a + \eta_a)r_a(\boldsymbol{\mu}) + \Phi_{\mathcal{A}\setminus\mathcal{I}}(\boldsymbol{\mu} + \boldsymbol{\eta}) \\
< \sum_{a\in\mathcal{I}}(\mu_a + \eta_a)r_a(\boldsymbol{\mu} + \boldsymbol{\eta}) + \Phi_{\mathcal{A}\setminus\mathcal{I}}(\boldsymbol{\mu} + \boldsymbol{\eta}),
\end{aligned}
$$

which again contradicts the hypothesis. $\quad\square$

Lemma A.2. *Let $\mathcal{P}_{\mathcal{A},\mathcal{H}}^{\mathcal{R}}$ be a Pareto front, and let $\boldsymbol{r}(\boldsymbol{\mu}) \in \mathcal{P}_{\mathcal{A},\mathcal{H}}^{\mathcal{R}}$ denote the solution the linear weighting Problem given by Eq.(2). Let $\mathcal{I} \subseteq \mathcal{A}$ and let $\boldsymbol{\mu}^{\mathcal{I}} : \mu_{\mathcal{I}}^{\mathcal{I}} > 0, \mu_{\mathcal{A}\setminus\mathcal{I}}^{\mathcal{I}} = 0, \|\boldsymbol{\mu}^{\mathcal{I}}\|_1^1 = 1$. Then*

$$(\exists i \in \mathcal{I} : r_i(\boldsymbol{\mu}^{\mathcal{I}}) < \|\boldsymbol{r}(\boldsymbol{\mu})\|_\infty) \vee (\boldsymbol{\mu} \in \boldsymbol{\mu}^*).$$

Proof. By contradiction, assume

$$(r_i(\boldsymbol{\mu}^{\mathcal{I}}) \geq \|\boldsymbol{r}(\boldsymbol{\mu})\|_\infty, \forall i \in \mathcal{I}) \wedge (\boldsymbol{\mu} \notin \boldsymbol{\mu}^*).$$

We can then write

$$
\begin{aligned}
\sum_{a\in\mathcal{A}}\mu_a^{\mathcal{I}}r_a(\boldsymbol{\mu}^{\mathcal{I}}) &\geq \sum_{a\in\mathcal{A}}\mu_a^{\mathcal{I}}\|\boldsymbol{r}(\boldsymbol{\mu})\|_\infty \\
&> \sum_{a\in\mathcal{A}}\mu_a^{\mathcal{I}}\|\boldsymbol{r}(\boldsymbol{\mu}^*)\|_\infty \\
&\geq \sum_{a\in\mathcal{A}}\mu_a^{\mathcal{I}}r_a(\boldsymbol{\mu}^*),
\end{aligned}
$$

which contradicts the definition of $\boldsymbol{r}(\boldsymbol{\mu}^{\mathcal{I}})$. $\quad\square$

Theorem 5.1. Let $\mathcal{P}_{\mathcal{A},\mathcal{H}}^{\mathcal{R}}$ be a Pareto front, and $\boldsymbol{r}(\boldsymbol{\mu}) \in \mathcal{P}_{\mathcal{A},\mathcal{H}}^{\mathcal{R}}$ denote the solution to the linear weighting Problem 2. For any $\boldsymbol{\mu}' \notin \underset{\boldsymbol{\mu}\in\Delta^{|\mathcal{A}|-1}}{\arg\min} \|\boldsymbol{r}(\boldsymbol{\mu})\|_\infty$, and $\boldsymbol{\mu}^* \in \underset{\boldsymbol{\mu}\in\Delta^{|\mathcal{A}|-1}}{\arg\min} \|\boldsymbol{r}(\boldsymbol{\mu})\|_\infty$, the sets

$$N_i = \{\boldsymbol{\mu} : r_i(\boldsymbol{\mu}) < \|\boldsymbol{r}(\boldsymbol{\mu}')\|_\infty\}$$

satisfy:

1. $\boldsymbol{\mu}^* \in \bigcap\limits_{i\in\mathcal{A}} N_i$;

2. If $\boldsymbol{\mu} \in N_i \to \lambda\boldsymbol{\mu} + (1-\lambda)\boldsymbol{e}^i \in N_i, \forall \lambda \in [0,1], i = 1,\ldots,|\mathcal{A}|$, where $\boldsymbol{e}^i$ denotes the standard basis vector;

3. $\forall \mathcal{I} \subseteq \mathcal{A}, \boldsymbol{\mu} : \mu_{\mathcal{A}\setminus\mathcal{I}} = 0 \to \boldsymbol{\mu} \in \bigcup\limits_{i\in\mathcal{I}} N_i$;

4. If $\boldsymbol{r}(\boldsymbol{\mu})$ is also continuous in $\boldsymbol{\mu}$, then $\forall \mathcal{I} \subseteq \mathcal{A}$ such that $\boldsymbol{\mu} \in \bigcap\limits_{i\in\mathcal{I}} N_i \to \exists \epsilon > 0 : B_\epsilon(\boldsymbol{\mu}) \subset \bigcap\limits_{i\in\mathcal{I}} N_i$;

5. If $\mathcal{P}_{\mathcal{A},\mathcal{H}}^{\mathcal{R}}$ is also convex, then $\boldsymbol{r}(\boldsymbol{\mu}^*) \in \underset{\boldsymbol{r}\in\mathcal{P}_{\mathcal{A},\mathcal{H}}^{\mathcal{R}}}{\arg\min} \|\boldsymbol{r}\|_\infty$.

Proof.

Property 1: We can directly observe that

$$r_a(\boldsymbol{\mu}^*) \leq \|\boldsymbol{r}(\boldsymbol{\mu}^*)\|_\infty < \|\boldsymbol{r}(\boldsymbol{\mu}')\|_\infty, \forall a \in \mathcal{A},$$

and therefore $\boldsymbol{\mu}^* \in \bigcap\limits_{i\in\mathcal{A}} N_i$.

Property 2: Direct application of Lemma A.1.

Property 3: Direct application of Lemma A.2.

Property 4: For all $i \in \mathcal{I}$ we have $r_i(\boldsymbol{\mu}) < \|\boldsymbol{r}(\boldsymbol{\mu}')\|_\infty$, since $r_i(\boldsymbol{\mu})$ is also continuous in $\boldsymbol{\mu}$, $\exists \epsilon_i > 0 : \forall \boldsymbol{\mu}'' \in B_{\epsilon_i}(\boldsymbol{\mu}) \to r_i(\boldsymbol{\mu}) < \|\boldsymbol{r}(\boldsymbol{\mu}')\|_\infty$. Taking $\epsilon = \min\limits_{i\in\mathcal{I}} \epsilon_i$ we have $B_\epsilon(\boldsymbol{\mu}) \subset \bigcap\limits_{i\in\mathcal{I}} N_i$.

Property 5: Immediate since every point in the Pareto front in this condition can be expressed as a solution to the linear weighting Problem 2. Proven in Theorem 4.1 and (Geoffrion, 1968).

$\quad\square$

A.2. Analysis of Proposed Optimization Method

Here we discuss some of the properties of the APStar Algorithm (Algorithm 1). Key observations regarding the update sequence can be summarized as follows:

- Updates $\boldsymbol{\mu}^{t+1} = (\boldsymbol{\mu}^t + \frac{1}{K\|\mathbf{1}_{\mu^t}\|_1^1}\mathbf{1}_{\mu^t})\frac{K}{K+1}$ always satisfy $\boldsymbol{\mu}^t \geq 0, \|\boldsymbol{\mu}^t\|_1^1 = 1$,

- Consecutive updates that do not decrease the minimax risk have a step size that converges to 0: $\|\boldsymbol{\mu}^{t+1} - \boldsymbol{\mu}^t\|_2^2 = \frac{1}{K+1}\|\boldsymbol{\mu}^t - \mathbf{1}_{\mu^t}\|_2^2 \leq \frac{1}{K+1} \to 0$,

- The update sequence $\boldsymbol{\mu}^{t+1} = (\boldsymbol{\mu}^t + \boldsymbol{\eta})\frac{K}{K+1}; K \leftarrow K+1$ converges to $\boldsymbol{\eta}$. That is $\boldsymbol{\mu}^{t+1} \to \boldsymbol{\eta}$.

So far, we showed that the choice of update sequence always proposes feasible weighting vectors, with progressively smaller step sizes, but that nonetheless can reach any point in the feasible region given sufficient updates.

We can also state that the update directions $\frac{\mathbf{1}_\mu}{\|\mathbf{1}_\mu\|_1^1}$ are not fixed points of the algorithm unless they themselves are a viable update vector.

Lemma A.3. *Let $\boldsymbol{\mu}'$ with corresponding $\|r(\boldsymbol{\mu}')\|_\infty$ in the conditions of Theorem 5.1. Denote the possible update directions of Algorithm 1 as*

$$\boldsymbol{\eta}^{\mathcal{I}} : \eta_i = \begin{cases} \frac{1}{|\mathcal{I}|}, & i \in \mathcal{I}, \\ 0 & o.w. \end{cases}$$

Similarly, denote

$$\mathbf{1}^{\mu} : \mathbf{1}_i^{\mu} = \begin{cases} 1 \ if \ r_i(\boldsymbol{\mu}) > \|r(\boldsymbol{\mu}')\|_\infty, \\ 0 \ o.w. \end{cases}$$

All update directions that are non-viable descent updates are repellors. That is, if $\mathcal{I} \subseteq \mathcal{A} : \|r(\boldsymbol{\eta}^{\mathcal{I}})\|_\infty > \|r(\boldsymbol{\mu}')\|_\infty$ then $\exists \epsilon : \forall \boldsymbol{\mu} \in B_\epsilon(\boldsymbol{\eta}^{\mathcal{I}}); \frac{\mathbf{1}_\mu}{\|\mathbf{1}_\mu\|_1^1} \neq \boldsymbol{\eta}^{\mathcal{I}}.$

Proof. This is a direct corollary of properties 3 and 4 of Theorem 5.1 □

A full convergence proof of the algorithm would need to show that this algorithm has no cycles. This can be motivated to a point by observing that the update directions are equivalent to performing gradient descent on the following function:

$$\begin{aligned} F(\boldsymbol{\mu}) &= \sum_{i=1}^{|\mathcal{A}|} (1 - \mu_a)\mathbf{1}(r_i(\boldsymbol{\mu}) \geq \bar{\boldsymbol{r}}), \\ \nabla F(\boldsymbol{\mu}) &= -\{(r_i(\boldsymbol{\mu}) \geq \bar{\boldsymbol{r}})\}_{i=1}^{|\mathcal{A}|} = -\mathbf{1}_\mu, \end{aligned} \quad (3)$$

where we abuse the notation $\nabla F(\boldsymbol{\mu})$ since the function is not differentiable on the boundaries where $r_i(\boldsymbol{\mu}) = \bar{r}$. We do note however that even on those points, $\mathbf{1}_\mu$ is a valid descent direction. Function $F(\boldsymbol{\mu})$ can be shown to have no non-global minima in the set $\boldsymbol{\mu} : \boldsymbol{\mu} > 0, \|\boldsymbol{\mu}\|_1^1 = 1$, and global minina on all $\boldsymbol{\mu} \in \bigcap_{i \in \mathcal{A}} N_i$. We do note, however, that gradient descent on discontinous functions can still produce cycles on pathological cases.

Figure 4 (repeating Figure 3 for convenience) shows examples of iterations of the APStar algorithm across randomly generated star-convex sets satisfying the conditions of Theorem 5.1. We observe that the algorithm converges to a viable update direction in all instances. Convergence for simple cases happens in few iterations, but challenging scenarios can still be appropriately solved with the proposed algorithm; on average, our algorithm is significantly faster than random sampling, especially in challenging scenarios. An explanation on how these distributions are sampled is presented next.

SAMPLING STAR-CONVEX DISTRIBUTIONS

Here we describe a simple procedure to sample Star-convex distributions in Δ^2 that satisfy the properties of Theorem 5.1.

We note that any point in the Δ^2 simplex can be embedded into $\mathbb{R}^2$ by using the following function,

$$\begin{aligned} f(\boldsymbol{\mu}) &: \quad \Delta^2 \to \mathbb{R}^2, \\ f(\boldsymbol{\mu}) &= \quad (\tfrac{2\mu_1 + \mu_2}{2}, \tfrac{\sqrt{3}\mu_2}{2}). \end{aligned}$$
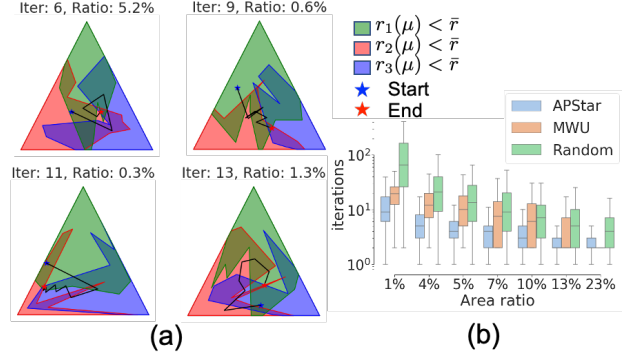

Figure 4. Synthetic data experiment on star-shaped sets. (a) Randomly sampled star sets satisfying the conditions of Theorem 5.1; a random starting point is sampled (Blue), and the trajectories recovered by the APStar algorithm are recorded until convergence (Red); number of iterations and intersection area are shown for all examples. (b) Empirical distribution of the number of iterations required to converge as a function of the percentage of linear weights that lie in the triple intersection; values are shown for the APStar algorithm, random sampling, and the multiplicative weight update (MWU) algorithm proposed in (Chen et al., 2017) for minimax optimization. The number of iterations required by the algorithm is well below both samplers, this is especially apparent for low area ratio scenarios. APStar finds a viable weight in all scenarios, with simpler sets and initial conditions requiring a smaller number of iterations on average.

In this restricted $\mathbb{R}^2$ space, we note parametric curves of the form $\mathcal{C}_i : [0, \frac{\pi}{3}] \to \mathbb{R}^+$ can be used to parametrize the Star-shaped sets we require for Theorem 5.1. Namely, for any Star-shaped set $N_i \in \Delta^2$ centered on $\boldsymbol{e}^i$, we can find a function $\mathcal{C}_i : [0, \frac{\pi}{3}] \to \mathbb{R}^+$ such that

$$\begin{aligned} N_i = \{\boldsymbol{\mu} \in \Delta^2 : d_i = \langle f(\boldsymbol{\mu}) - f(\boldsymbol{e}^i), f(\boldsymbol{e}^{i+1}) - f(\boldsymbol{e}^i)\rangle, \\ \theta_i = \angle(f(\boldsymbol{\mu}) - f(\boldsymbol{e}^i), f(\boldsymbol{e}^{i+1}) - f(\boldsymbol{e}^i), \\ d_i < \mathcal{C}_i(\theta_i)\} \bigcup \{f(\boldsymbol{e}^i)\}. \end{aligned}$$

Figure 5 illustrates the relationship between curves $\mathcal{C}_i$ and their corresponding Star-shaped sets N_i.

To create a curve $\mathcal{C}_i$, we construct a piecewise-linear function by sampling K tuples $(\theta_i^j, r_i^j)_{j=0}^K$ satisfying

$$\begin{aligned} 0 &= \theta_i^0 < \cdots < \theta_i^K = \pi/3, \\ r_i^j &\in [0, 1] \ \forall j = 0, \ldots, K. \end{aligned}$$

In our convergence experiments, we set $K = 7$, $r_i^j \sim U[0, 1]$ and $(\theta_i^j)_{j=1}^{K-1} = \text{Sort}((u \sim U[0, \pi/3])_{j=1}^{K-1})$. We sample $\mathcal{C}_0, \mathcal{C}_1, \mathcal{C}_2$ independently and then reject these functions if the corresponding sets N_0, N_1, N_2 do not satisfy the properties of Theorem 5.1. Namely, properties 2 and
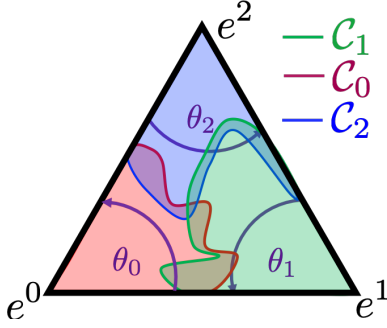
Figure 5. Illustration of Star-shaped sets $N_i \in \Delta^2$. The sets N_i can be parametrized by the boundary curves $\mathcal{C}_i$, all points that connect the boundary $\mathcal{C}_i$ with $f(e^i)$ conform the Star-shaped set N_i.

4 are satisfied by construction, we check that Property 1 is satisfied by verifying that $N_0 \cup N_1 \cup N_2$ contains at least one element.

Property 3 is verified by checking that $N_0 \cap N_1 \cap N_2$ covers the entire triangle in $\mathcal{R}^2$, and also that

$$\mathcal{C}_0(0) + \mathcal{C}_1(\pi/3) > 1,$$
$$\mathcal{C}_1(0) + \mathcal{C}_2(\pi/3) > 1,$$
$$\mathcal{C}_2(0) + \mathcal{C}_0(\pi/3) > 1.$$

A.3. MMPF Implementation Details

Here we present implementation details to estimate the Minimax Pareto Fair classifier from data. As mentioned in Section 5, the APStar algorithm (Algorithm 1) requires an optimizer to solve the linear weighting problem (Problem 2). We propose two options, one minimizes it using stochastic gradient descent (SGD), we call this approach joint estimation, and it is described in Algorithm 2. Note that each batch samples the sensitive attributes uniformly in order to reduce the variance of the conditional risk estimators for every group in every batch. The second approach was also presented in Section 5 and is called plug-in. Here each conditional distribution $(p(Y|X, A), p(A|X))$ is estimated independently from the data, and the optimal model for a given weighting vector $\boldsymbol{\mu}$ is computed using the expression derived in Theorem 4.2, Algorithm 3 describes this approach.

In both the joint and plug-in approach, the APStar algorithm makes decision based on the risk vector evaluated on the validation datasets. This is done to empirically improve generalization and disallow excessive overfitting.

Algorithm 2 Joint Estimation

Input: Train $D^{tr} = \{(x^i, y^i, a^i)\}_{i=1}^{N_{tr}}$, Validation $D^{val} = \{(x^i, y^i, a^i)\}_{i=1}^{N_{val}}$, Network: h_{θ^o}, Weights: $\boldsymbol{\mu}$, Loss $\ell(\cdot, \cdot)$, Learning Rate: lr, Decay rate $= \gamma$, Epochs: n_E, Batch Size: n_B, Maximum Patience $= n_P$
$h^* \leftarrow h_{\theta^o}$, epochs$\leftarrow 0$, patience $\leftarrow 0$
repeat
    $t \leftarrow 0, h_\theta \leftarrow h^*$
    while $t < \frac{N_{tr}}{n_B}$ **do**
        $\{a_i\}_{i=1}^{n_B} \sim U[1, .., |\mathcal{A}|]; \{x_i, y_i | a_i\}_{i=1}^{n_B} \sim D^{tr}$
        $\boldsymbol{r}(h_\theta) \leftarrow \left\{ \frac{\sum_{i=1}^{n_B} \mathbf{1}[a_i=a]\ell(h_\theta(x_i), y_i)}{\sum_{i=1}^{n_B} \mathbf{1}[a_i=a]} \right\}_{a \in \mathcal{A}}$
        $\theta \leftarrow \theta - lr \nabla_\theta \langle \boldsymbol{\mu}, \boldsymbol{r}(h_\theta) \rangle$
        $t \leftarrow t + 1$
    end while
    epochs += 1 # *epoch ended; evaluate on validation*
    $\boldsymbol{r}^{val}(h_\theta) \leftarrow \left\{ \frac{\sum_{i \in D^{val}} \mathbf{1}[a_i=a]\ell(h_\theta(x_i), y_i)}{\sum_{i \in D^{val}} \mathbf{1}[a_i=a]} \right\}_{a \in \mathcal{A}}$
    if $\langle \boldsymbol{\mu}, \boldsymbol{r}^{val}(h_\theta) \rangle \leq \langle \boldsymbol{\mu}, \boldsymbol{r}^{val}(h^*) \rangle$ **then**
        $h^* \leftarrow h_\theta$; patience $\leftarrow 0$
    else
        $lr \leftarrow \gamma lr$; patience += 1
    end if
until epochs$\geq n_E \vee$ patience $\geq n_P$
return $h^*, \boldsymbol{r}^{val}(h^*)$

Algorithm 3 Plug-in Estimation

Input: Validation $D^{val} = \{(x^i, y^i, a^i)\}_{i=1}^{N_{val}}$, Networks estimating $p(Y|X, a)$: $\{p_{\theta_a}^Y\}_{a=1}^{|\mathcal{A}|}$ and $p(A|X)$: p_ϕ^A, Priors $\{p_a\}_{a=1}^{|\mathcal{A}|}$, Weights: $\boldsymbol{\mu}$, Loss $\ell(\cdot, \cdot)$.
$h^*(x) \leftarrow \frac{\sum_{a=1}^{|\mathcal{A}|} p_{\theta_a}^Y(x) p_\phi^A(x) \frac{\mu_a}{p_a}}{\sum_{a=1}^{|\mathcal{A}|} p_\phi^A(x) \frac{\mu_a}{p_a}}$
$\boldsymbol{r}^{val}(h^*) \leftarrow \left\{ \frac{\sum_{i \in D^{val}} \mathbf{1}[a_i=a]\ell(h^*(x_i), y_i)}{\sum_{i \in D^{val}} \mathbf{1}[a_i=a]} \right\}_{a \in \mathcal{A}}$
return $h^*, \boldsymbol{r}^{val}(h^*)$

A.4. Synthetic Data Experiments

We tested our approach on synthetic data where the observations are drawn from the following distributions:

$$\begin{aligned}
A &\sim U[1, ..., |\mathcal{A}|], \\
X|A = a &\sim N(m_a, 1), \\
Y|X = x, A = a &\sim Ber(f_a(x)), \\
f_a(x) &= \rho_a^l \mathbf{1}[x \leq t_a] + \rho_a^h \mathbf{1}[x > t_a].
\end{aligned} \tag{4}$$

Note that $f_a(x)$ is a piecewise-constant function. We used Brier Score as our loss function, then the Bayes-optimal classifier for each group is $f_a(x)$, while the optimal classifier for the linear weighting problem can be computed numerically using the expression derived in Theorem 4.2.

For the synthetic experiments presented in this section we chose $|A| = 3$, $\{m_0, m_1, m_2\} = \{-0.5, 0, 0.5\}$, $\{t_0, t_1, t_2\} = \{-0.25, 0, 0.25\}$, $\rho_{0,1,2}^l = 0.1$, $\rho_{0,1}^h = 0.9$, and $\rho_2^h = 0.8$.

Figure 6.a shows the conditional distributions $p(X|a)$ and $p(Y|X,a)$ obtained with these parameters, along with the minimax Pareto fair classifier.

We evaluate performance of the APStar algorithm performs when $h^{\mu}, r(\mu)$ are computed using the closed form formula in Theorem 4.2, expectations are computed via numerical integration. This enables us to evaluate the performance of the algorithm in the infinite sample and perfect h optimization regime. We recover the optimal minimax Pareto fair weights μ^* via grid search, since a closed form solution for these weights is not available. Figure 6.b and Figure 6.c show how the risk vector $r(\mu)$ and linear weights μ approach the minimax optimal r^*, μ^* as a function of iterations of the APStar algorithm. Figure 2 shown in Section 5 was also generated with these parameters.

Figures 6.d and 6.e show the performance of the algorithm as a function of training samples, the classifier is implemented using an NN, and is minimized using SGD. Table 7 in Section A.9 provides the architecture and optimization details. We observe that the optimal classifier is non-linear (see Figure 6.a) which motivates the use of NNs for estimation. Note that relative errors decrease with the number of samples. In both cases, the algorithm is able to effectively converge to the minimax Pareto fair risk.

The Pareto curve shown in figure 1 (Section 3) was generated with parameters $|A| = 2$, $\{m_0, m_1\} = \{-0.5, 0.5\}$, $\{t_0, t_1\} = \{-0.25, 0.25\}$, $\{\rho_0^l, \rho_1^l\} = \{0.3, 0.05\}$, $\{\rho_0^h, \rho_1^h\} = \{0.7, 0.95\}$ and Brier Score risk.
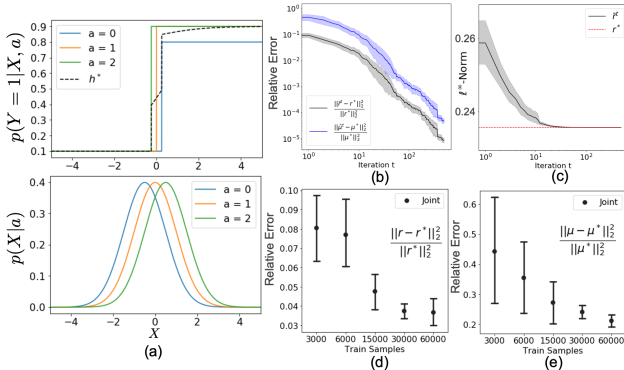

Figure 6. Synthetic data experiment. (a) Conditional distributions $p(X|a)$ and $p(Y = 1|X,a)$, the minimax Pareto fair classifier h^* is also shown. (b) The APStar algorithm converges to the optimal risk and weight vectors in a scenario where access to ground truth joint distributions is provided. (c) shows how the minimax risk is reduced in this scenario as well. (d) and (e) show minimax convergence of risk and weight vectors respectively as a function of samples when the classifier h^{μ} is estimated with a fully connected neural network. Relative error quickly decays when more samples are acquired.

A.5. Plug in vs Joint Estimation Analisis

We empirically compare the performance of the plug-in and joint estimation approaches presented in Section 4. The main advantage of plug-in estimation is that once the conditional classifiers are calculated, evaluating a new weight vector does not require any optimization; in contrast, joint estimation requires a full optimization run for each new weight vector. The main advantage of joint estimation is that it requires a single model, and makes use of all samples to train it; this can be beneficial when samples are scarce, and can be motivated by the transfer learning literature. In our problem setting, we can consider sensitive groups as different domains or tasks, and our goal is to find a model that has the best minimax performance. If the conditional distributions $p(Y|X,a)$ and $p(X|a)$ match for every $a \in \mathcal{A}$ ($Y \perp A|X$ and $X \perp A$), we are in the optimal transfer learning scenario where all groups benefit from each others' samples to estimate the target. In this situation, the joint approach would be expected to perform better on the test set than the plug-in approach. From this ideal case we can identify the following two deviations:

- Case I: $Y \perp A|X$ and $d(p(X|a), p(X|a')) \geq \epsilon$, $\forall a, a' \in \mathcal{A}, \epsilon > 0$;

- Case II: $X \perp A$ and $d(p(Y|X,a), p(Y|X,a')) \geq \epsilon$, $\forall a, a' \in \mathcal{A}, \epsilon > 0$;

where $d(.,.)$ is some distance or divergence between distributions.

Case I keeps $Y \perp A|X$, but $p(X|a)$ differ across a values. It is reasonable to assume that as ϵ increases, the difference in the group conditional risks on test data for joint and plug-in estimation will be low. Note that in this scenario, if the hypothesis class is unbounded, there is no trade-off between group risks as shown in Lemma 4.3, hence the Pareto front is the Utopia point; we expect the joint estimation approach to outperform plug-in estimation at the same number of weight updates.

Case II keeps $X \perp A$, but $p(Y|X,a)$ differ across a values. In this scenario, there may be a trade-offs between group risks. We argue that in a finite sample scenario, as ϵ increases, it is not clear if joint estimation will be better or worse than plug-in since the former may be affected by negative transfer.

EXPERIMENTS

We empirically examine these cases by choosing $|\mathcal{A}| = 2$ and simulating synthetic data from the following distribu-

tion:

$$X|0 \sim \mathcal{N}(0,1), \qquad X|1 \sim \mathcal{N}(m_1, 1), m_1 \geq 0,$$
$$Y|X, 0 \sim Ber(f_0(X)), \quad Y|X, 1 \sim Ber(f_1(X)),$$
$$A \sim Ber(\tfrac{1}{2}),$$

(5)

where

$$f_0(X) = (0.6 + 0.2\mathbf{1}[x \geq 0])sign[sin(2\pi X) + 1]$$
$$+0.2 - 0.1\mathbf{1}[x \geq 0],$$
$$f_1(X) = (1 - \lambda)f_0(X) + \lambda[1 - round(f_0(X))],$$
$$\lambda \in [0, 1].$$

(6)

Note that parameter m_1 is used to change the separation between $p(X|0)$ and $p(X|1)$; we measure this using KL-divergence $D_{KL}(p(X|0), p(X|1)) = \frac{m_1^2}{2}$. The parameter λ controls the difference between $p(Y|X, 1)$ and $p(Y|X, 0)$, we measure this using $E_X[D_{KL}(p(Y|X, 0), p(Y|X, 1))]$, which can be numerically approximated. For Case I we choose $\lambda = 0$ and $m_0 \in \{0, 0.5, 1, 1.5, 2\}$; for Case II we choose $m_0 = 0$ and $\lambda \in \{0, 0.2, 0.5, 0.8\}$.

Figure 7 shows three examples of synthetic data generated with these distributions were the input variable X, sensitive variable A, and target Y exhibit various dependencies. Figures 7.a and 7.b show examples of conditional distributions $p(X|A)$ and $p(Y = 1|X, A)$ for Cases I and II. Figure 7.c shows an example were both $X \not\perp A$ and $Y \not\perp A$. Additionally, we plot the MMPF classifier $h^*(X)$. When $Y \perp A|X$ we have that $h^*(X) = p(Y|X)$ and when $X \perp A$, $h^*(X)$ is a linear combination of $p(Y|X, 0)$ and $p(Y|X, 1)$ with the minimax weight coefficients. These are direct consequences of Theorem 4.2. In Figure 7.c $h^*(X)$ shows a more complex interaction since both $p(X|a)$ and $p(Y|X, a)$ differ between a values. Asymptotically, since $p(a = 1|X = \infty) = 1, p(a = 0|X = -\infty) = 1$, we recover $h^*(\infty) = p(Y|X, 1)$ and $h^*(-\infty) = p(Y|X, 0)$.
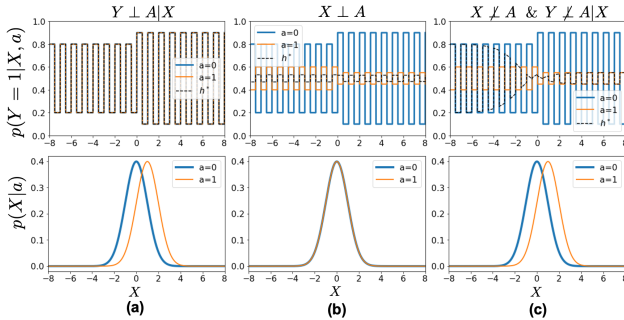
Figure 7. (a), (b), and (c) show conditional distributions for $p(Y = 1|X, a)$ (top) and $p(X|a)$ (bottom), simulated according to Eq.5 and Eq.6 for three situations. (a) corresponds to case I ($Y \perp A|X$), (b) to case II ($X \perp A$) and (c) to $Y \not\perp A|X$ and $X \not\perp A$. All plots of $p(Y = 1|X, a)$ include the optimal minimax Pareto fair classifier $h^*(X)$.

Figure 8 compares the performance of the plug-in and joint estimation approach. On both approaches, we limit the number of weight updates of the APStar algorithm to 15. Figure 8.a compares relative differences in the risk vector as a function of the divergence between $p(X|0)$ and $p(X|1)$ for $4k$ and $9k$ train samples and Case I ($Y \perp A$). Figure 8.b compares accuracies under the same conditions. We observe that in this scenario, the benefit of joint estimation is evident, especially with a small number of samples. Note that this gap is reduced as $D_{KL}(p(X|0)||p(X|1))$ and the number of samples increases; as expected, both methods perform better when more samples are available.

Figures 8.c and 8.d show the same comparisons for Case II. In this scenario there is no clear difference between plug-in and joint estimation, though the former appears to be marginally better; both methods improve performance with additional samples.
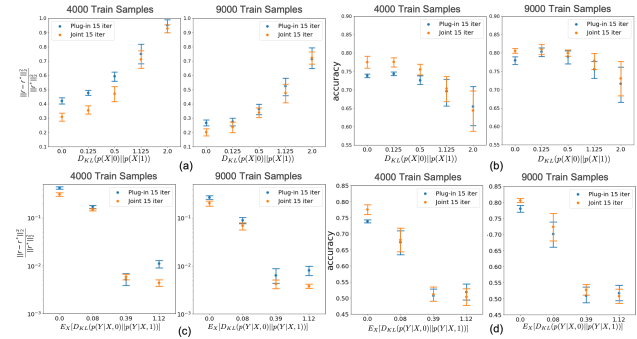
Figure 8. (a) Relative error of the estimated Minimax Pareto Fair risk as a function of divergence between $p(X|0)$ and $p(X|1)$ for Case I ($Y \perp A|X$) at $4k$ and $9k$ training samples. (b) shows the corresponding accuracy comparison. (c) and (d) mirror (a) and (b) for Case II ($X \perp A$).

UNBALANCED CLASSIFICATION: $Y = A$

In balanced risk minimization, we have $Y = A$. If the risk function is either Cross Entropy or Brier Score, we can apply Theorem 4.1 and recover

$$h^{\boldsymbol{\mu}}(x) = \left\{ \frac{\frac{\mu_a}{p(a)}p(a|x)}{\sum_{a' \in \mathcal{A}} \frac{\mu'_a}{p(a')}p(a'|x)} \right\}_{a=1}^{|\mathcal{A}|}.$$

This particular scenario is noteworthy because the plug-in approach only needs to estimate $p(a|x) = p(y|x)$, all Pareto optimal classifiers can be easily derived from $p(y|x)$ by simply re-weighting each component of the output probability vector. This re-weighting requires no expensive minimization procedure, and enables extensive iterations of the APStar algorithm to find the optimal weight vector $\boldsymbol{\mu}$. For these types of scenarios, it is advantageous to estimate $p(y|x)$ by using a Naive or Balanced classifier, and then derive all

optimal classifiers via a simple weighting of the output vector, the optimal weights can still be found using our APStar algorithm.

OBSERVATION SUMMARY

We summarize our observations from these experiments and discussions.

- Joint estimation may benefit from transfer learning and seems to be no worse than plug-in estimation even when target conditional distributions do not match. Note that a certain amount of negative transfer may be required by the minimax Pareto Fair classifier, which may negate the advantages of plug-in estimation when target conditional distributions differ.

- Plug-in estimation requires multiple ($|\mathcal{A}| + 1$) models, while joint estimation only requires one; this makes this approach impractical in some scenarios. However, the former approach allows for cheap iterations of the APStar algorithm.

- In the balanced classification problem ($Y = A$), plug-in estimation requires the same number of models as joint estimation, but is cheaper to evaluate.

A.6. Methods

We compare the performance of the following methods:

Kamishima. (Kamishima et al., 2012) uses logistic regression as a baseline classifier, and requires numerical input (observations), and binary target variable. Fairness is controlled via a regularization term with a tuning parameter η that controls the trade-off between fairness and overall accuracy. η is optimized via grid search with $\eta \in (0, 300)$ as in the original paper. We report results on the hyperparameter configuration that produces the best minimax cross-entropy across sensitive groups.

Feldman. (Feldman et al., 2015) provides a preprocessing algorithm to sanitize input observations. It modifies each input attribute so that the marginal distribution of each coordinate is independent of the sensitive attribute. The degree to which these marginal distributions match is controlled by a λ parameter between 0 and 1. It can handle numerical and categorical observations, as well as non-binary sensitive attributes, and arbitrary target variables. Following (Friedler et al., 2019), we train a linear logistic regressor on top of the sanitized attributes. λ is optimized via grid search with increments of 0.05. We report results on the hyperparameter configuration that produces the best minimax cross-entropy across sensitive groups.

Zafar. (Zafar et al., 2015) Addresses disparate mistreatment via a convex relaxation. Specifically, in the implementation provided in (Friedler et al., 2019), they train a logistic regression classifier with a fairness constraint that minimizes the covariance between the sensitive attribute and the classifier decision boundary. This algorithm can handle categorical sensitive attributes and binary target variables, and numerical observations. The maximum admissible covariance is handled by a hyperparameter c, tuned by logarithmic grid search with values between 0.001 and 1. We report results on the hyperparameter configuration that produces the best minimax cross-entropy across sensitive groups.

Hardt. (Hardt et al., 2016) proposes a post-processing algorithm that takes in an arbitrary predictor and the sensitive attribute as input, and produces a new, fair predictor that satisfies equalized odds. This algorithm can handle binary target variables, an arbitrary number of sensitive attributes, and any baseline predictor, but requires test-time access to sensitive attributes. it does not contain any tuning parameter. We apply this method on top of both the Naive Classifier and our Pareto Fair classifier.

Naive Classifier (Naive). Standard classifier, trained to minimize an expected risk $h = \arg\min_{h \in \mathcal{H}} E_{X,A,Y}[\ell(h(X), \delta^Y)]$. The baseline classifier class $\mathcal{H}$ is implemented as a neural network and varies by experiment as described in Section A.9, the loss function also varies by experiment and is also described in Section A.9. Optimization is done via stochastic gradient descent.

Balanced Classifier (Balanced). Baseline classifier designed to address undersampling of minority classes, trained to mimimize a class-rebalanced expected risk $h = \arg\min_{h \in \mathcal{H}} E_{A \sim U[1,...,|\mathcal{A}|], (X,Y) \sim P(X,Y|A)}[\ell(h(X), \delta^Y)]$. Like the Naive classifier, it is implemented as a neural network and optimized via stochastic gradient descent. The sole difference with the Naive classifier is that, during training, samples are drawn from the new input distribution $A \sim U[1, \ldots, |\mathcal{A}|]$; $X, Y|A \sim P(X, Y|A)$, which is achieved by re-weighted sampling of the original training dataset.

Minimax Pareto Fair (MMPF, joint and plug-in). Our proposed methodology, finds a Pareto-optimal model h^* such that it has minimax Pareto risk $\mathbf{r}^*$. It achieves this by searching for weighting coefficients μ^* such that $\mathbf{r}^*$ is the solution to the corresponding linear weighted problem (see Eq.2). This method alternates between minimizing a linearly-weighted loss function (Eq.2), and updating the weighting coefficients according to Algorithm 1. The baseline classifier class $\mathcal{H}$ is implemented as a neural network and varies by experiment as described in Section A.9, the loss function also varies by experiment and is also described

in Section A.9. The minimization of the weighted loss function at every step of the APStar algorithm is implemented via stochastic gradient descent (joint estimation) or by direct application of the closed form optimal classifier derived in Theorem 4.2 (plug-in estimation). The latter approach requires estimating $p(a|x)$, $p(y|x,a)$ and $p(a)$.

A.7. Evaluation Metrics

Here we describe the metrics used to evaluate the performance of all tested methods. We are given a set of test samples $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^N$ where $x_i \in \mathcal{X}$ is a realization of our model input and $y_i \in \mathcal{Y}$ the corresponding objective. We assume that $\mathcal{Y}$ is a finite alphabet, as in a classification problem; we will represent the one-hot encoding of y_i as δ^{y_i}. Given a trained model $h : \mathcal{X} \rightarrow [0,1]^{|\mathcal{Y}|}$ the predicted output for an input x_i is a vector $\boldsymbol{h}(x_i) : \|\boldsymbol{h}(x_i)\|_1^1 = 1$ (e.g., output of a softmax layer). The predicted class is $\hat{y}_i = \arg\max_j h_j(x_i)$ and its associated confidence is $\hat{p}_i = \max_j h_j(x_i)$. Ideally $\hat{y}_i$ should be the same as y_i. Using these definitions, we compute the following metrics.

Accuracy (Acc): $\frac{1}{N} \sum_{i=1}^N \mathbf{1}(y_i = \hat{y}_i)$. Fraction of correct classifications in dataset.

Brier Score (BS): $\frac{1}{N} \sum_{i=1}^N \|\delta^{y_i} - \vec{p}_i\|^2$ where δ^{y_i} is the one-hot representation of the categorical ground truth value y_i. This quantity is also known as Mean Square Error (MSE).

Cross Entropy (CE): $-\frac{1}{N} \sum_{i=1}^N \langle \delta^{y_i}, \log \boldsymbol{h}(x_i) \rangle$ also known as negative log-likelihood (NLL) of the multinomial distribution.

Expected Calibration Error (ECE): $\frac{1}{N} \sum_{m=1}^M \left| \sum_{i \in B_m} [\mathbf{1}(y_i = \hat{y}_i) - \hat{p}_i] \right|$ where M is the number of bins to divide the interval $[0,1]$ such that $B_m = \{i \in \{1,..,N\} : \hat{p}_i \in (\frac{m-1}{M}, \frac{m}{M}]\}$ are the group of samples that our model assigns a confidence ($\hat{p}_i$) in the interval $(\frac{m-1}{M}, \frac{m}{M}]$. Measures how closely the predicted probabilities match the true base rates.

Maximum Calibration Error (MCE): $max_{m \in \{1,...,M\}} \left| \frac{1}{|B_m|} \sum_{i \in B_m} [\mathbf{1}(y_i = \hat{y}_i) - \hat{p}_i] \right|$. Measures worst-case miscalibration errors.

These metrics are computed independently for each sensitive subgroup on the test set and reported in Section A.10.

A.8. Details on Experiments on Real Data

The following is a description of the data and experiments for each of the real datasets. The information present here is summarized in Table 6.

MIMIC-III. This dataset consist of clinical records collected from adult ICU patients at the Beth Israel Deaconess Medical Center (MIMIC-III dataset) (Johnson et al.,

Table 6. Basic characteristics of real datasets

Dataset	Objective	Sensitive Attribute	Train/ Val/ Test	Splits
Adult (Dua & Graff, 2017a)	2 categories: Income	4 categories: Gender (F/M), Ethnicity(W/NW)	60/20/20	5
German (Dua & Graff, 2017b)	2 categories: Credit	2 categories: Gender (F/M)	60/20/20	5
MIMIC-III (Johnson et al., 2016)	2 categories: Mortality (A/D)	8 categories: Mortality(A/D), Age (A/S) Ethnicity (W/NW)	60/20/20	5
HAM10000 (Tschandl et al., 2018)	7 categories: Type of lesion	7 categories: Type of lesion	60/20/20	5

2016). The goal is predicting patient mortality from clinical notes. We follow the pre-processing methodology outlined in (Chen et al., 2018), where we analyze clinical notes acquired during the first 48 hours of ICU admission; discharge notes were excluded, as where ICU stays under 48 hours. Tf-idf statistics on the $10,000$ most frequent words in clinical notes are taken as input features.

We identify 8 sensitive groups as the combination of age (under/over 55 years old), ethnicity as determined by the majority group (white/nonwhite); and outcome (alive/deceased). Here we will use the term adult to refer to people under 55 years old and senior otherwise. This dataset shows large sample disparities since 56.7% corresponds to the overall majority group (alive-senior-white) and only 0.4% to the overall minority group (deceased-adult-nonwhite).

We used a fully connected neural network as described in Table 7 as the baseline classifier for our proposed MMPF framework. We compare our results against both the Naive and Naive Balanced algorithms using the same neural network architecture, and use crossentropy (CE) as our training loss. We also evaluate the performance of Zafar, Feldman and Kamishima applied on the feature embeddings learned by the Naive classifier (results over the original input features failed to converge on the provided implementations).

We report the performance across a 5-fold split of the data, we used a 60/20/20 train-validation-test partition as described on Table 6 and report results over the test set. We denote the overall sensitive attribute as the combination of outcome (A:alive/D:deceased), age (A:adult/S:senior) and ethnicity (W:white, NW:nonwhite) with shorthand notation of the form D/A/W to denote, for example, deceased, white adult. We also note that results on Zafar, Kamishima and Hardt were done over only the sensitive attributes Adult/Senior and White/Nonwhite, outcome was not considered as a sensitive attribute for both methods. This was done because Hardt requires test-time access to sensitive attributes, which would not be possible for the outcome variable, and Zafar attempts to decorrelate sensitive attributes and classification decision boundaries, which is counterproductive when the sensitive attribute includes the correct

decision outcome.

HAM10000. This dataset contains over $10,000$ dermatoscopic images of skin lesions over a diverse population (Tschandl et al., 2018). Lesions are classified in 7 diagnostic categories, and the goal is to learn a model capable of identifying the category from the lesion image. The dataset is highly unbalanced since 67% of the samples correspond to a melanocytic nevi lesion (nv), and 1.1% to dermatofibroma (df).

Here we chose to use the diagnosis class as both the target and sensitive variable, casting balanced risk minimization as a particular use-case for the proposed Pareto fairness framework.

We load a pre-trained DenseNet121 network (Huang et al., 2017) and train it to classify skin lesions from dermatoscopic images using our Pareto fairness framework. We compared against the Naive and the Balanced training setup. Note that in the Balanced approach we use a batch sampler where images from each class have the same probability, this can be seen as a naive oversampling technique. Table 7 shows implementation details.

We report the performance across 5-fold split of the data, we used a 60/20/20 train-validation-test partition, and report results over the test set. For each group we follow the original notation: Actinic keratoses and intraepithelial carcinoma / Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv) and vascular lesions (vasc).

Adult. The Adult UCI dataset (Dua & Graff, 2017a) is based on the 1994 U.S. Census and contains data on $32,561$ adults. The data contains 105 binarized observations representing education status, age, ethnicity, gender, and marital status, and a target variable indicating income status (binary attribute representing over or under $50,000$). Following (Friedler et al., 2019), we take ethnicity and gender as our target sensitive attributes, defining four subgroups (White/Other and Male/Female). We also present results considering just the gender as sensitive attribute (Male/Female). To compare our MMPF framework evenly against the other methods, we limit our hypothesis class to linear logistic regression (MMPF LR). Additionally, we also show results for a Neural Network model (MMPF); a model class that satisfies the convexity property.

German. The German credit dataset (Dua & Graff, 2017a) contains 20 observations collected across 1000 individuals, and a binary target variable assessing the individual's credit score as good or bad. We consider gender (Male/Female) as the sensitive attribute, which is not included in the data but can be inferred. As in the Adult dataset, we limit our hypothesis class to linear logistic regression to compare evenly across methodologies.

A.9. Neural Network Architectures and Parameters

Table 7. Summary of network architectures and losses. All networks have a softmax output layer. ADAM was used as the training optimizer, with the specified learning rates (lr) and batch size n_B. Logistic Regression was trained using the implementation provided in Sklearn (Pedregosa et al., 2011).

Dataset	Network Body	Gate	Loss type	Parameters training
Synthetic	Dense ResNet (512x512)x2	ELU	BS	n_B=512 lr=1e-3
Adult German	Logistic Regression (LR)	-	CE	-
Adult	Dense ResNet (512x512)x2	-	CE	n_B=32 lr=5e-4
MIMIC-III	FullyConnected 2048x2048	ELU	CE/BS	n_B=512 lr=1e-6/5e-6
HAM10000	DenseNet121 (Huang et al., 2017)	ReLU	BS	n_B=32 lr=5e-6

Table 7 summarizes network architectures and loss functions for all experiments in this paper (Section 6 and supplementary material). Note that all networks have a standard dense softmax as their final layer. The training optimizer is ADAM (Kingma & Ba, 2014), loss functions were either crossentropy (CE) or Brier Score (BS), also known as categorical mean square error (MSE).

For joint estimation, the weights $\boldsymbol{\mu}$ were initialized uniformly, we selected maximum patience $n_P = 20$, a decay rate $\gamma = 0.25$ and a maximum of 500 epochs. All experiments terminated from lack of generalization improvement rather than maximum number of epochs. For the APStar algorithm we picked $\alpha = 0.5$, we allowed a maximum number of 20 iterations. The regularization parameter in the Sklearn implementation of logistic regression was set to $C = 1e6$ following (Friedler et al., 2019).

In the plug-in approach, each conditional probability was estimated with the architectures and parameters specified in Table 7 maximum patience $n_P = 20$, a decay rate $\gamma = 0.25$ and a maximum of 500 epochs. Here we allowed the APStar algorithm to have a maximum of 500 iterations since each iteration does not require optimization, only risk evaluation.

A.10. Supplementary Results

Here we present expanded results on all datasets. Accuracy (Acc), Brier Score (BS), Cross-Entropy (CE), Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) are displayed per sensitive group. Mean and standard deviations are reported when avaialble across 5 splits. Disparity between best and worst groups is computed per split, and the mean and standard deviation of this value is reported. Note that this way of computing disparity may lead to seemingly large disparity values, since the worst and best performing group per split may differ.

Table 8. HAM10000 dataset. We underline the worst group metric per method, and bold the one with the best minimax performance. Smallest disparity is also bolded.

	Acc comparison							
Group	akiec	bcc	bkl	df	mel	nv	vasc	Disparity
Ratio	3.3%	5.1%	11%	1.1%	11.1%	67%	1.4%	65.9%
Naive	39.1±5.7 %	58.4±7.0 %	51.5±4.3 %	2.6±3.5%	43.7±4.3 %	93.7±1.1 %	66.9±6.6 %	91.1±3.9%
Balanced	67.9±6.9 %	73.4±4.2 %	58.2±9.9 %	75.7±7.6 %	58.1±5.8 %	73.5±1.6 %	83.9±8.3 %	32.5±4.6%
MMPF P	64.2±5.1 %	66.2±5.8 %	**63.9±6.8** %	69.6±13.5%	67.5±3.9%	64.1±1.0%	71.2±1.0%	**19.8±6.6%**

	BS comparison							
Group	akiec	bcc	bkl	df	mel	nv	vasc	Disparity
Ratio	3.3%	5.1%	11%	1.1%	11.1%	67%	1.4%	65.9%
Naive	0.816±0.082	0.586±0.083	0.675±0.068	1.384±0.043	0.808±0.043	0.093±0.015	0.48±0.102	1.291±0.042
Balanced	0.459±0.089	0.37±0.048	0.579±0.1	0.392±0.106	0.565±0.069	0.361±0.02	0.211±0.106	0.45±0.066
MMPF P	0.494±0.078	0.463±0.065	0.49±0.074	0.447±0.131	0.447±0.042	**0.5±0.015**	0.38±0.127	**0.228±0.058**

	CE comparison							
Group	akiec	bcc	bkl	df	mel	nv	vasc	Disparity
Ratio	3.3%	5.1%	11%	1.1%	11.1%	67%	1.4%	65.9%
Naive	1.924±0.321	1.19±0.188	1.405±0.142	4.178±0.209	1.589±0.104	0.195±0.029	1.069±0.282	3.983±0.19
Balanced	0.944±0.172	0.715±0.096	1.199±0.211	0.898±0.295	1.122±0.128	0.787±0.038	0.456±0.281	0.95±0.218
MMPF P	1.011±0.177	0.913±0.158	0.949±0.135	1.047±0.401	0.85±0.068	**1.128±0.033**	0.804±0.242	**0.615±0.156**

	ECE comparison							
Group	akiec	bcc	bkl	df	mel	nv	vasc	Disparity
Ratio	3.3%	5.1%	11%	1.1%	11.1%	67%	1.4%	65.9%
Naive	0.211±0.019	0.125±0.026	0.189±0.032	0.598±0.047	0.287±0.03	0.03±0.001	0.156±0.049	0.568±0.047
Balanced	0.139±0.046	0.078±0.013	0.135±0.063	**0.183±0.035**	0.119±0.035	0.066±0.014	0.119±0.028	**0.143±0.023**
MMPF P	0.133±0.028	0.113±0.029	0.107±0.036	0.213±0.049	0.082±0.023	0.135±0.01	0.133±0.035	0.151±0.034

	MCE comparison							
Group	akiec	bcc	bkl	df	mel	nv	vasc	Disparity
Ratio	3.3%	5.1%	11%	1.1%	11.1%	67%	1.4%	65.9%
Naive	0.616±0.262	0.353±0.124	0.534±0.151	0.962±0.018	0.555±0.091	0.315±0.219	0.521±0.156	0.744±0.082
Balanced	0.505±0.157	0.383±0.177	0.49±0.22	0.548±0.145	0.272±0.024	0.227±0.06	0.636±0.126	0.474±0.089
MMPF P	0.369±0.134	0.362±0.156	0.362±0.172	**0.59±0.098**	0.266±0.032	0.285±0.037	0.556±0.107	**0.444±0.058**

Table 9. MIMIC dataset. We underline the worst group metric per method, and bold the one with the best minimax performance. Smallest disparity is also bolded. Standard deviations are computed across 5 splits.

Acc comparison

Group	A/A/NW	A/A/W	A/S/NW	A/S/W	D/A/NW	D/A/W	D/S/NW	D/S/W	Disparity
Ratio	5.7%	13.3%	12.9%	56.7%	0.4%	0.9%	1.8%	8.3%	56.3%
Naive CE	98.7±0.8%	98.8±0.5%	97.6±0.5%	98.0±0.3%	26.0±10.0%	34.5±3.9%	20.6±2.1%	22.6±2.6%	79.4±1.7%
Naive BS	99.0±0.4%	98.8±0.4%	97.8±0.5%	98.1±0.3%	26.7±9.3%	34.6±4.4%	19.0±2.0%	21.2±1.9%	80.5±1.3%
Balanced CE	85.8±1.9%	86.2±1.1%	76.4±1.8%	79.2±0.4%	76.1±8.5%	80.1±3.3%	66.9±2.4%	67.3±2.1%	22.4±2.8%
Balanced BS	87.9±1.2%	87.5±1.4%	77.5±2.1%	79.3±0.5%	74.4±7.1%	78.8±3.4%	66.8±2.2%	68.0±2.1%	22.6±2.3%
Zafar	91.9±1.5%	93.9±1.4%	91.6±0.8%	93.2±0.5%	49.2±9.4%	41.7±9.2%	33.2±4.1%	32.0±2.4%	62.9±3.6%
Feldman	97.4±1.9%	97.7±1.9%	95.2±3.4%	95.6±3.3%	33.0±16.4%	35.4±8.0%	28.7±2.4%	31.9±3.6%	72.1±5.5%
Kamishima	98.5±1.2%	98.4±0.7%	96.8±0.9%	96.7±0.7%	26.7±9.8%	37.5±5.0%	25.1±5.1%	29.6±2.8%	76.4±5.2%
MMPF CE	83.3±2.1%	83.1±1.0%	71.3±1.4%	74.0±1.2%	81.6±6.5%	83.2±3.9%	73.2±3.0%	74.7±2.9%	**16.2±2.8%**
MMPF BS	86.0±0.9%	84.8±1.3%	**72.6±1.7%**	74.1±0.5%	78.9±9.5%	81.7±3.8%	73.1±2.0%	75.2±2.1%	17.1±3.5%
MMPF CE P	82.4±3.2%	81.8±1.2%	70.7±1.7%	74.4±0.8%	80.6±7.7%	84.5±3.0%	73.0±3.5%	72.4±3.1%	17.4±2.5%
MMPF BS P	81.8±5.6%	81.8±2.6%	70.7±2.1%	74.9±1.0%	78.0±8.6%	81.9±2.5%	72.6±2.8%	72.8±3.5%	17.8±3.8%
Naive BS+H	59.8±1.8%	59.9±1.9%	59.7±2.1%	59.8±2.1%	53.2±6.4%	51.2±3.2%	50.3±2.0%	50.5±0.9%	12.3±3.8%
Balanced BS+H	76.7±1.3%	77.1±1.1%	76.9±2.4%	76.7±1.4%	67.3±10.3%	66.7±2.9%	65.6±2.8%	66.6±3.2%	19.1±1.8%
Zafar+H	64.0±2.2%	64.2±1.4%	64.3±2.3%	64.2±1.9%	53.4±9.2%	52.9±4.5%	51.5±2.8%	52.1±1.6%	17.8±3.1%
Feldman+H	61.1±2.5%	61.2±2.4%	61.2±2.4%	61.3±2.5%	55.0±11.4%	49.7±1.9%	52.1±2.9%	50.8±3.1%	16.9±7.4%
Kamishima+H	60.0±0.8%	60.2±1.5%	60.2±1.3%	60.2±1.4%	53.9±6.7%	52.4±1.7%	51.6±2.2%	52.3±1.5%	11.9±1.9%
MMPF BS+H	72.6±2.3%	72.7±1.1%	72.1±1.7%	72.5±1.3%	72.1±8.6%	72.1±2.8%	**72.0±3.7%**	72.2±2.6%	**11.4±3.5%**

BS comparison

Group	A/A/NW	A/A/W	A/S/NW	A/S/W	D/A/NW	D/A/W	D/S/NW	D/S/W	Disparity
Ratio	5.7%	13.3%	12.9%	56.7%	0.4%	0.9%	1.8%	8.3%	56.3%
Naive CE	0.029±0.004	0.029±0.005	0.054±0.007	0.048±0.003	0.995±0.141	0.896±0.072	1.086±0.014	1.053±0.022	1.064±0.019
Naive BS	0.034±0.004	0.035±0.005	0.059±0.007	0.053±0.003	0.961±0.134	0.836±0.068	1.051±0.013	1.022±0.016	1.025±0.021
Balanced CE	0.2±0.025	0.198±0.014	0.313±0.017	0.284±0.005	0.369±0.067	0.292±0.043	0.417±0.018	0.421±0.026	0.246±0.039
Balanced BS	0.19±0.018	0.188±0.013	0.307±0.019	0.283±0.004	0.386±0.062	0.31±0.039	0.418±0.014	0.413±0.026	0.251±0.04
Zafar	0.157±0.03	0.118±0.028	0.162±0.015	0.131±0.009	1.009±0.203	1.143±0.177	1.323±0.082	1.344±0.05	1.25±0.072
Feldman	0.038±0.026	0.035±0.03	0.072±0.051	0.068±0.052	1.172±0.303	1.116±0.104	1.258±0.077	1.213±0.067	1.288±0.084
Kamishima	0.027±0.011	0.026±0.008	0.063±0.011	0.06±0.008	1.053±0.163	0.936±0.095	1.062±0.041	0.993±0.033	1.108±0.078
MMPF CE	0.237±0.031	0.236±0.019	0.372±0.014	0.342±0.011	0.303±0.063	0.233±0.037	0.346±0.016	0.344±0.03	0.167±0.028
MMPF BS	0.212±0.015	0.217±0.017	**0.354±0.016**	0.337±0.006	0.331±0.057	0.254±0.035	0.352±0.017	0.338±0.028	0.17±0.027
MMPF CE P	0.274±0.033	0.269±0.017	0.373±0.013	0.343±0.011	0.325±0.061	0.271±0.025	0.363±0.012	0.37±0.027	**0.144±0.017**
MMPF BS P	0.272±0.06	0.262±0.03	0.368±0.014	0.332±0.01	0.323±0.079	0.256±0.032	0.357±0.022	0.365±0.032	0.168±0.042

CE comparison

Group	A/A/NW	A/A/W	A/S/NW	A/S/W	D/A/NW	D/A/W	D/S/NW	D/S/W	Disparity
Ratio	5.7%	13.3%	12.9%	56.7%	0.4%	0.9%	1.8%	8.3%	56.3%
Naive CE	0.073±0.003	0.072±0.005	0.122±0.009	0.112±0.003	1.536±0.175	1.314±0.11	1.561±0.036	1.53±0.038	1.549±0.027
Naive BS	0.093±0.006	0.093±0.009	0.139±0.01	0.129±0.004	1.407±0.15	1.179±0.103	1.458±0.041	1.425±0.029	1.394±0.049
Balanced CE	0.332±0.036	0.324±0.02	0.473±0.021	0.435±0.008	0.575±0.092	0.458±0.051	0.605±0.026	0.613±0.037	0.329±0.052
Balanced BS	0.322±0.026	0.317±0.017	0.467±0.024	0.437±0.007	0.593±0.079	0.481±0.044	0.604±0.022	0.601±0.034	0.332±0.043
Zafar	1.71±0.468	1.199±0.298	1.834±0.261	1.363±0.127	14.611±4.841	17.197±4.31	21.829±2.593	22.653±1.918	21.729±2.164
Feldman	0.072±0.047	0.064±0.054	0.138±0.1	0.127±0.095	3.924±0.962	3.475±0.612	3.802±0.842	3.854±0.74	4.211±0.75
Kamishima	0.057±0.015	0.054±0.011	0.128±0.011	0.121±0.011	1.838±0.176	1.561±0.167	1.585±0.077	1.521±0.049	1.808±0.155
MMPF CE	0.378±0.041	0.373±0.026	0.547±0.019	0.508±0.012	0.501±0.101	0.377±0.045	0.517±0.023	0.517±0.04	0.232±0.037
MMPF BS	0.349±0.022	0.352±0.023	0.524±0.021	0.503±0.008	0.532±0.07	0.407±0.04	0.525±0.025	0.509±0.037	0.231±0.036
MMPF CE P	0.438±0.039	0.430±0.024	0.550±0.014	0.516±0.015	0.505±0.078	0.438±0.029	0.542±0.016	0.551±0.03	**0.17±0.023**
MMPF BS P	0.431±0.072	0.416±0.038	0.541±0.015	0.498±0.015	0.51±0.104	0.412±0.042	0.534±0.026	0.544±0.038	0.212±0.053

ECE comparison

Group	A/A/NW	A/A/W	A/S/NW	A/S/W	D/A/NW	D/A/W	D/S/NW	D/S/W	Disparity
Ratio	5.7%	13.3%	12.9%	56.7%	0.4%	0.9%	1.8%	8.3%	56.3%
Naive CE	0.049±0.006	0.046±0.003	0.071±0.007	0.068±0.004	0.519±0.105	0.434±0.065	0.574±0.014	0.542±0.025	0.537±0.011
Naive BS	0.068±0.003	0.065±0.006	0.088±0.004	0.087±0.004	0.502±0.106	0.386±0.051	0.576±0.014	0.54±0.018	0.519±0.005
Balanced CE	0.066±0.015	0.049±0.011	0.039±0.013	0.029±0.008	0.173±0.048	0.1±0.03	0.073±0.029	0.057±0.022	0.148±0.042
Balanced BS	0.083±0.004	0.073±0.011	0.042±0.011	0.043±0.009	0.151±0.043	0.079±0.041	0.069±0.023	0.044±0.009	**0.126±0.044**
Zafar	0.079±0.015	0.059±0.014	0.081±0.008	0.065±0.004	0.511±0.099	0.576±0.089	0.666±0.039	0.671±0.025	0.627±0.037
Feldman	0.015±0.009	0.012±0.011	0.023±0.016	0.019±0.015	0.624±0.152	0.565±0.071	0.628±0.044	0.587±0.034	0.668±0.06
Kamishima	0.029±0.005	0.021±0.004	0.055±0.011	0.05±0.007	0.543±0.112	0.44±0.069	0.547±0.037	0.486±0.021	0.568±0.06
MMPF CE	0.041±0.015	0.029±0.015	0.055±0.018	0.03±0.008	0.152±0.057	0.08±0.021	0.076±0.03	0.039±0.013	0.143±0.04
MMPF BS	0.064±0.009	0.047±0.011	0.037±0.011	0.021±0.006	0.178±0.056	0.091±0.034	0.073±0.016	0.038±0.013	0.167±0.044
MMPF CE P	0.105±0.024	0.086±0.016	0.042±0.016	0.049±0.015	0.177±0.063	0.152±0.016	0.102±0.018	0.059±0.027	0.159±0.044
MMPF BS P	0.077±0.033	0.052±0.02	0.041±0.021	0.025±0.007	**0.144±0.037**	0.122±0.027	0.065±0.017	0.044±0.026	0.127±0.027

MCE comparison

Group	A/A/NW	A/A/W	A/S/NW	A/S/W	D/A/NW	D/A/W	D/S/NW	D/S/W	Disparity
Ratio	5.7%	13.3%	12.9%	56.7%	0.4%	0.9%	1.8%	8.3%	56.3%
Naive CE	0.277±0.094	0.231±0.034	0.207±0.027	0.206±0.011	0.835±0.119	0.878±0.071	0.839±0.05	0.874±0.046	0.729±0.031
Naive BS	0.315±0.075	0.262±0.046	0.247±0.039	0.23±0.022	0.806±0.122	0.869±0.063	0.822±0.057	0.861±0.058	0.707±0.028
Balanced CE	0.138±0.024	0.087±0.034	0.074±0.025	0.049±0.013	0.383±0.072	0.225±0.167	0.191±0.077	0.119±0.043	0.393±0.062
Balanced BS	0.155±0.039	0.114±0.034	0.077±0.018	0.072±0.016	0.395±0.29	0.14±0.045	0.159±0.085	0.109±0.047	0.36±0.269
Zafar	0.483±0.227	0.647±0.15	0.591±0.192	0.445±0.086	0.619±0.122	0.678±0.164	0.696±0.059	0.696±0.034	0.495±0.154
Feldman	0.272±0.094	0.283±0.222	0.245±0.108	0.219±0.129	0.864±0.068	0.705±0.052	0.736±0.05	0.734±0.06	0.738±0.142
Kamishima	0.416±0.056	0.158±0.026	0.191±0.019	0.154±0.025	0.823±0.133	0.747±0.162	0.787±0.08	0.785±0.054	0.73±0.067
MMPF CE	0.102±0.043	0.059±0.022	0.097±0.028	0.068±0.02	0.346±0.119	0.24±0.156	0.147±0.05	0.073±0.027	0.352±0.114
MMPF BS	0.111±0.024	0.076±0.024	0.076±0.031	0.041±0.012	0.347±0.072	0.25±0.133	0.149±0.039	0.063±0.013	0.343±0.069
MMPF CE P	0.163±0.044	0.128±0.014	0.078±0.03	0.077±0.014	0.489±0.216	0.249±0.053	0.177±0.045	0.126±0.035	0.424±0.219
MMPF BS P	0.132±0.047	0.091±0.028	0.078±0.026	0.044±0.012	**0.327±0.101**	0.253±0.152	0.12±0.03	0.083±0.033	**0.317±0.125**

Table 10. Adult ethnicity and gender dataset. We underline the worst group metric per method, and bold the one with the best minimax performance. Smallest disparity is also bolded. Standard deviations are computed across 5 splits.

Acc comparison

type	Female Other	Male Other	Female White	Male White	disc
Ratio	6.0%	7.7%	26.1%	60.3%	54.3%
Naive LR	94.7±0.9%	84.0±1.0%	91.8±0.4%	<u>80.6±0.5%</u>	14.1±1.0%
Balanced LR	95.0±1.0%	84.5±0.7%	91.9±0.4%	<u>80.5±0.5%</u>	14.5±1.0%
Zafar	95.1±0.9%	84.1±1.4%	92.0±0.2%	<u>80.6±0.5%</u>	14.5±0.9%
Feldman	95.1±1.0%	83.7±1.3%	91.8±0.4%	<u>80.4±0.3%</u>	14.7±0.9%
Kamishima	95.3±1.0%	83.8±0.6%	91.9±0.4%	<u>80.0±1.2%</u>	15.2±1.8%
MMPF LR	94.6±0.7%	84.7±1.0%	91.3±0.3%	<u>80.6±0.5%</u>	14.0±1.0%
MMPF LR P	94.6±0.7%	84.0±1.0%	91.4±0.6%	<u>80.7±0.5%</u>	13.9±1.0%
MMPF	94.6±1.2%	84.4±0.9%	91.5±0.5%	<u>80.9±0.6%</u>	13.6±1.5%
MMPF P	94.5±1.1%	84.3±1.5%	90.7±0.5%	**81.0±0.8%**	**13.4±1.5%**
Naive LR+H	76.3±1.0%	72.6±2.2%	75.6±1.6%	<u>71.7±2.0%</u>	5.6±0.8%
Balanced LR+H	76.5±1.2%	72.3±2.2%	75.7±2.0%	<u>71.6±2.1%</u>	5.7±0.6%
Zafar+H	73.8±2.5%	70.4±2.3%	73.3±2.5%	<u>69.7±3.0%</u>	5.0±0.7%
Feldman+H	74.8±2.7%	71.2±2.6%	74.2±2.3%	<u>70.7±2.6%</u>	5.5±1.1%
Kamishima+H	72.3±3.5%	68.5±3.2%	71.4±3.6%	<u>68.0±4.0%</u>	**4.8±0.7%**
MMPF LR+H	79.2±1.4%	75.4±1.5%	78.2±0.6%	**74.4±1.8%**	5.5±1.3%

BS comparison

type	Female Other	Male Other	Female White	Male White	disc
Ratio	6.0%	7.7%	26.1%	60.3%	54.3%
Naive LR	0.079±0.009	0.207±0.012	0.119±0.004	<u>0.266±0.005</u>	0.187±0.011
Balanced LR	0.078±0.01	0.207±0.011	0.119±0.004	<u>0.267±0.006</u>	0.189±0.011
Zafar	0.076±0.01	0.208±0.015	0.118±0.003	<u>0.265±0.005</u>	0.189±0.01
Feldman	0.082±0.011	0.213±0.014	0.122±0.003	<u>0.269±0.004</u>	0.187±0.011
Kamishima	0.081±0.01	0.216±0.008	0.119±0.005	<u>0.271±0.016</u>	0.191±0.021
MMPF LR	0.084±0.009	0.206±0.013	0.126±0.005	<u>0.264±0.005</u>	0.18±0.01
MMPF LR P	0.085±0.007	0.209±0.014	0.127±0.006	<u>0.264±0.005</u>	0.18±0.007
MMPF	0.084±0.012	0.209±0.009	0.126±0.005	**0.261±0.005**	0.177±0.014
MMPF P	0.086±0.013	0.209±0.014	0.137±0.007	<u>0.262±0.006</u>	**0.176±0.015**

CE comparison

type	Female Other	Male Other	Female White	Male White	disc
Ratio	6.0%	7.7%	26.1%	60.3%	54.3%
Naive LR	0.14±0.013	0.321±0.017	0.204±0.005	<u>0.408±0.008</u>	0.268±0.016
Balanced LR	0.138±0.013	0.322±0.018	0.203±0.004	<u>0.411±0.008</u>	0.273±0.015
Zafar	0.143±0.024	0.336±0.027	0.204±0.003	<u>0.409±0.01</u>	0.266±0.03
Feldman	0.149±0.015	0.332±0.019	0.21±0.005	<u>0.412±0.006</u>	0.262±0.016
Kamishima	0.146±0.014	0.337±0.013	0.202±0.006	<u>0.414±0.023</u>	0.269±0.026
MMPF LR	0.153±0.012	0.322±0.018	0.218±0.005	**0.404±0.007**	**0.251±0.015**
MMPF LR P	0.153±0.01	0.324±0.019	0.218±0.008	<u>0.405±0.006</u>	0.251±0.01
MMPF	0.141±0.017	0.326±0.015	0.219±0.009	<u>0.404±0.009</u>	0.263±0.022
MMPF P	0.151±0.021	0.331±0.029	0.245±0.007	<u>0.41±0.014</u>	0.258±0.03

ECE comparison

type	Female Other	Male Other	Female White	Male White	disc
Ratio	6.0%	7.7%	26.1%	60.3%	54.3%
Naive LR	0.02±0.006	<u>0.03±0.003</u>	0.019±0.003	0.014±0.004	0.018±0.003
Balanced LR	0.017±0.004	**0.025±0.005**	0.014±0.004	0.02±0.004	**0.014±0.006**
Zafar	0.013±0.003	<u>0.032±0.008</u>	0.015±0.003	0.012±0.004	0.021±0.011
Feldman	0.029±0.004	<u>0.034±0.009</u>	0.024±0.001	0.011±0.003	0.026±0.009
Kamishima	0.014±0.003	<u>0.028±0.005</u>	0.009±0.003	0.013±0.004	0.019±0.006
MMPF LR	0.036±0.007	<u>0.028±0.008</u>	0.036±0.006	0.01±0.003	0.031±0.005
MMPF LR P	<u>0.029±0.007</u>	0.033±0.009	0.024±0.009	0.013±0.002	0.025±0.003
MMPF	0.022±0.004	<u>0.031±0.003</u>	0.015±0.004	0.016±0.004	0.018±0.007
MMPF P	0.028±0.005	<u>0.033±0.007</u>	0.018±0.003	0.015±0.004	0.021±0.006

MCE comparison

type	Female Other	Male Other	Female White	Male White	disc
Ratio	6.0%	7.7%	26.1%	60.3%	54.3%
Naive LR	<u>0.146±0.079</u>	0.111±0.032	0.082±0.028	0.036±0.014	0.132±0.06
Balanced LR	<u>0.215±0.075</u>	0.086±0.017	0.062±0.022	0.059±0.019	0.161±0.088
Zafar	<u>0.174±0.046</u>	0.118±0.03	0.061±0.038	0.033±0.013	0.145±0.044
Feldman	<u>0.226±0.124</u>	0.107±0.028	0.083±0.018	0.027±0.008	0.202±0.115
Kamishima	<u>0.283±0.126</u>	0.103±0.027	0.081±0.024	0.031±0.008	0.252±0.12
MMPF LR	**0.109±0.033**	0.077±0.019	0.085±0.02	0.03±0.008	**0.088±0.031**
MMPF LR P	0.22±0.078	0.118±0.063	0.059±0.021	0.038±0.005	0.184±0.079
MMPF	<u>0.202±0.051</u>	0.076±0.02	0.107±0.03	0.043±0.013	0.166±0.046
MMPF P	<u>0.17±0.078</u>	0.087±0.024	0.085±0.03	0.038±0.017	0.151±0.067

Table 11. Adult gender dataset. We underline the worst group metric per method, and bold the one with the best minimax performance. Smallest disparity is also bolded. Standard deviations are computed across 5 splits.

Adult gender Acc

	Female	Male	Disparity
Ratio	32.1%	67.9%	35.9%
Naive LR	92.3±0.4	80.5±0.4	11.9±0.7
Balanced LR	92.3±0.3	80.3±0.7	12.0±0.7
Zafar	92.5±0.3	80.9±0.3	11.6±0.4
Feldman	92.3±0.3	80.7±0.2	11.6±0.1
Kamishima	92.6±0.4	80.9±0.4	11.7±0.7
MMPF LR	91.9±0.4	81.0±0.4	10.9±0.7
MMPF	92.1±0.3	81.3±0.3	10.8±0.5
MMPF LR P	92.0±0.4	81.0±0.5	11.0±0.6
MMPF P	91.7±0.3	**81.5±0.5**	**10.1±0.5**
Feldman+H	72.3±2.5%	76.5±2.7%	4.1±0.8%
Kamishima+H	73.5±2.0%	77.8±1.2%	4.3±1.4%
Zafar+H	73.3±2.7%	77.3±2.5%	4.0±1.1%
Naive LR+H	74.2±2.7%	78.6±2.1%	4.4±0.7%
Balanced LR+H	73.8±3.1%	77.7±2.9%	3.9±0.9%
MMPF LR+H	**76.0±2.2%**	79.8±1.6%	**3.8±1.3%**

Adult gender CE

	Female	Male	Disparity
Ratio	32.1%	67.9%	35.9%
Naive LR	.204±.009	.411±.006	.207±.007
Balanced LR	.204±.011	.416±.011	.211±.005
Zafar	.202±.018	.398±.006	.195±.023
Feldman	.201±.004	.403±.004	.203±.006
Kamishima	.189±.006	**.395±.004**	.206±.007
MMPF LR	.204±.008	.395±.006	.19±.011
MMPF	.21±.019	.403±.025	.193±.013
MMPF LR P	.208±.008	.395±.005	.187±.01
MMPF P	.227±.019	.403±.023	**.176±.014**

Adult gender BS

	Female	Male	Disparity
Ratio	32.1%	67.9%	35.9%
Naive LR	.116±.004	.268±.004	.152±.005
Balanced LR	.117±.005	.272±.007	.155±.005
Zafar	.11±.004	.258±.003	.147±.005
Feldman	.115±.003	.263±.002	.148±.003
Kamishima	.11±.005	.258±.003	.147±.005
MMPF LR	.117±.006	.257±.004	.14±.007
MMPF	.117±.004	**.255±.004**	.138±.007
MMPF LR P	.119±.005	.258±.004	.138±.007
MMPF P	.127±.003	.256±.005	**.129±.003**

Adult gender ECE

	Female	Male	Disparity
Ratio	32.1%	67.9%	35.9%
Naive LR	.026±.008	.013±.004	.013±.004
Balanced LR	.023±.007	.014±.005	.01±.006
Zafar	**.01±.002**	.01±.003	**.003±.001**
Feldman	.026±.003	.01±.005	.016±.006
Kamishima	.012±.003	.012±.002	.003±.003
MMPF LR	.032±.005	.011±.003	.021±.005
MMPF	.009±.002	.015±.003	.006±.004
MMPF LR P	.028±.002	.009±.002	.019±.002
MMPF P	.02±.006	.015±.001	.006±.005

Adult gender MCE

	Female	Male	Disparity
Ratio	32.1%	67.9%	35.9%
Naive LR	.064±.012	.027±.013	.037±.019
Balanced LR	.065±.034	.031±.009	.034±.027
Zafar	.058±.013	.032±.01	.027±.008
Feldman	.071±.017	.024±.013	.047±.012
Kamishima	.073±.008	.031±.009	.042±.002
MMPF LR	.072±.017	.033±.006	.039±.021
MMPF	**.057±.021**	.031±.003	**.026±.019**
MMPF LR P	.064±.004	.03±.004	.034±.004
MMPF P	.085±.023	.047±.01	.039±.024

Table 12. German dataset. We underline the worst group metric per method, and bold the one with the best minimax performance. Smallest disparity is also bolded. Standard deviations are computed across 5 splits.

German Acc

	Female	Male	Disparity
Ratio	29.5%	70.5%	41.0%
Naive LR	70.7±7.3	71.2±4.5	8.8±4.7
Balanced LR	71.6±5.9	70.9±4.1	5.8±3.6
Zafar	73.0±5.6	71.0±3.5	5.8±3.5
Feldman	73.5±8.6	**71.9±4.3**	7.9±4.4
Kamishima	68.8±6.8	72.7±2.6	6.0±4.4
MMPF LR	72.5±5.5	71.6±2.8	5.0±2.6
MMPF LR P	70.7±4.5	71.5±3.6	**4.4±0.5**
Naive LR+H	57.5±1.7	57.8±1.8	5.7±3.6
Balanced LR+H	60.5±4.2	60.9±4.5	4.6±3.3
Feldman+H	61.6±4.7	62.2±5.0	7.1±3.9
Kamishima+H	61.7±4.0	61.3±4.2	4.5±2.2
Zafar+H	59.8±4.0	60.5±4.9	6.6±4.9
MMPF LR+H	**65.7±4.7**	65.9±4.7	**3.6±1.7**

German CE

	Female	Male	Disparity
Ratio	29.5%	70.5%	41.0%
Naive LR	.607±.1	.559±.069	.127±.064
Balanced LR	.594±.082	.568±.068	.096±.05
Zafar	.567±.09	.735±.205	.273±.151
Feldman	.564±.096	.551±.063	.091±.068
Kamishima	.62±.064	.545±.062	.075±.067
MMPF LR	.565±.04	.544±.046	**.048±.041**
MMPF LR P	**.563±.043**	.537±.051	.057±.034

German BS

	Female	Male	Disparity
Ratio	29.5%	70.5%	41.0%
Naive LR	.404±.077	.379±.05	.094±.043
Balanced LR	.393±.069	.38±.043	.071±.036
Zafar	.379±.07	.383±.052	.072±.05
Feldman	**.375±.079**	.371±.045	.07±.05
Kamishima	.413±.051	.368±.044	.047±.048
MMPF LR	.379±.038	.368±.039	**.044±.03**
MMPF LR P	.381±.039	.363±.041	.051±.025

German ECE

	Female	Male	Disparity
Ratio	29.5%	70.5%	41.0%
Naive LR	.136±.047	.099±.029	.051±.036
Balanced LR	.136±.032	.107±.032	**.038±.025**
Zafar	.117±.039	.097±.038	.047±.022
Feldman	.129±.045	.091±.04	.052±.043
Kamishima	.125±.067	.082±.029	.052±.04
MMPF LR	.096±.036	.046±.019	.056±.046
MMPF LR P	**.088±.039**	.049±.008	.047±.034

German MCE

	Female	Male	Disparity
Ratio	29.5%	70.5%	41.0%
Naive LR	.308±.11	.23±.023	.095±.091
Balanced LR	.322±.127	.216±.088	.106±.046
Zafar	.255±.135	.206±.08	.138±.072
Feldman	.285±.096	.166±.048	.137±.083
Kamishima	.212±.084	.157±.052	**.062±.062**
MMPF LR	.18±.067	.11±.046	.093±.075
MMPF LR P	**.172±.073**	.11±.029	.106±.046