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 algorithms 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.

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

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

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 $\boldsymbol{r}(h') \prec \boldsymbol{r}(h'')$. Similarily, 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, $h^* \in \arg\min_{h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}} \max_{a \in \mathcal{A}} r_a(h) = \arg\min_{h \in \mathcal{P}_{\mathcal{A},\mathcal{H}}} \|\boldsymbol{r}(h)\|_\infty$, 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*

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

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 $\mathcal{H} \subseteq \{h : \mathcal{X} \to [0,1]^{|\mathcal{Y}|}\}$ is convex iff $\forall\, h', h'' \in \mathcal{H},\ \lambda \in [0,1], \to \lambda h' + (1-\lambda)h'' \in \mathcal{H}$.

A risk function $r : \mathcal{H} \to \mathbb{R}_+$ is convex iff $\forall \lambda \in [0,1]\ r(\lambda h' + (1-\lambda)h'') \leq \lambda r(h') + (1-\lambda) r(h'')$

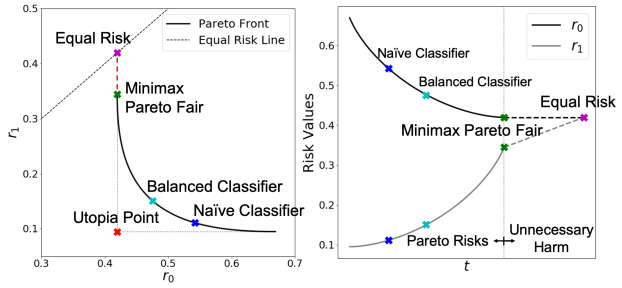

Figure 1. Achievable risk trade-offs for a binary classification problem $Y \in \{0,1\}$ with two unbalanced groups $A \in \{0,1\}$, and covariates $X|A \sim N(\mu_A, 1)$ (parameters provided in Supplementary Material). *Left: Pareto front risks and the equal risk line; the minimax Pareto fair point (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 $\forall \boldsymbol{r}, \boldsymbol{r}' \in \mathcal{P}^{\mathcal{R}}_{\mathcal{A},\mathcal{H}}, \lambda \in [0,1], \exists \boldsymbol{r}^{\lambda} \in \mathcal{P}^{\mathcal{R}}_{\mathcal{A},\mathcal{H}}$ such that $\boldsymbol{r}^{\lambda} \preceq \lambda \boldsymbol{r} + (1-\lambda)\boldsymbol{r}'$

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:

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

We use the shorthand notation $h^{\boldsymbol{\mu}}$ to describe a classifier that solves Problem 2; likewise, we denote $\boldsymbol{r}(\boldsymbol{\mu}) = \boldsymbol{r}_{h^{\boldsymbol{\mu}}}$. 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: $\forall \boldsymbol{r}, \boldsymbol{r}' \in \mathcal{P}^{\mathcal{R}}_{\mathcal{A},\mathcal{H}}, \lambda \in [0,1], \exists \boldsymbol{r}'' \in \mathcal{P}^{\mathcal{R}}_{\mathcal{A},\mathcal{H}} : \boldsymbol{r}'' \preceq \lambda \boldsymbol{r} + (1-\lambda)\boldsymbol{r}'$.
2. Every Pareto solution is a solution to Problem 2: $\forall \hat{\boldsymbol{r}} \in \mathcal{P}^{\mathcal{R}}_{\mathcal{A},\mathcal{H}}, \exists \boldsymbol{\mu} : \hat{\boldsymbol{r}} = \boldsymbol{r}(\boldsymbol{\mu})$.

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

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

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 $H(Y|X, A = a) = E_{X|A=a}[H(Y|X = X, A = a)]$.

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*

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

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

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 $\cup_{i \in \mathcal{I}} N_i$.

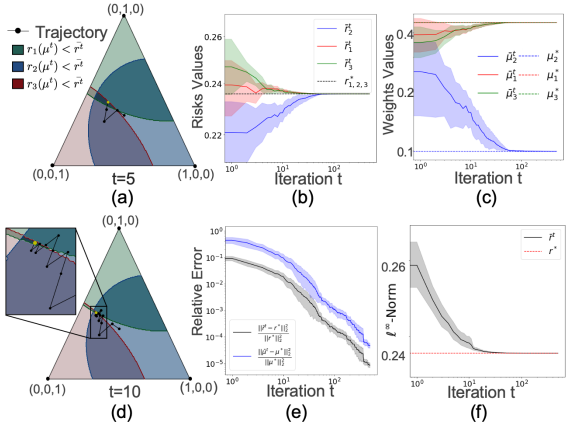

Figure 2. Synthetic data experiment with 3 sensitive groups. (a) and (d) show a simplex diagram of the linear weights $\boldsymbol{\mu}$ on the fifth and tenth iteration of the APStar algorithm; blue, green and red shaded areas correspond to the N_i areas at iterations, $t = 5, t = 10$, the optimal linear 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

  $\mathbf{1}_\mu \leftarrow \{\mathbf{1}(r_i(\mu) \geq \bar{\boldsymbol{r}})\}_{i=1}^{|\mathcal{A}|}$

  $\boldsymbol{\mu} \leftarrow (\alpha\boldsymbol{\mu} + \frac{1-\alpha}{K\|\mathbf{1}_\mu\|_1^1}\mathbf{1}_\mu)\frac{K}{(K-1)\alpha+1}$

  $h, \boldsymbol{r}(\mu) \leftarrow \{\arg\}\min_{h\in\mathcal{H}} \sum \mu_i r_i(h)$; $K \leftarrow K+1$

  if $\|\boldsymbol{r}(\mu)\|_\infty < \bar{\boldsymbol{r}}$ **then**

    $\bar{\boldsymbol{r}} \leftarrow \|\boldsymbol{r}(\mu)\|_\infty$, $K \leftarrow \min(K, K_{min})$

    $h^*, \boldsymbol{\mu}^*, \boldsymbol{r}^* \leftarrow h, \boldsymbol{\mu}, \boldsymbol{r}(\mu)$

  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 ($\boldsymbol{\mu}^{t+1} \rightarrow \alpha\boldsymbol{\mu}^t + (1-\alpha)\boldsymbol{\mu}^{\mathcal{A}\setminus\mathcal{I}}$). 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

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. 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.

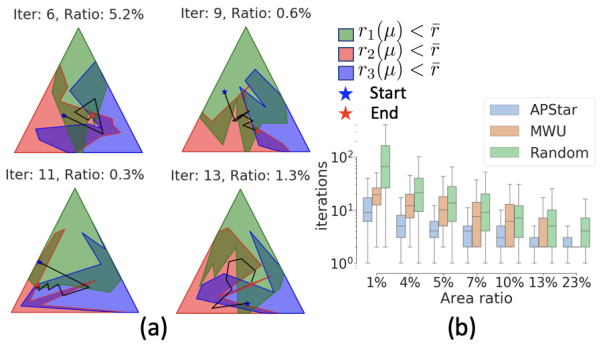

Figure 3. Synthetic data experiment on star-shaped sets. (a) Randomly sampled star sets satisfying the conditions of Theorem 5.1; a starting point is sampled (Blue), trajectories recovered by the APStar algorithm are recorded until convergence (Red); number of iterations and intersection area are shown 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 sensitive groups equally (Balanced). 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), 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, 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 exhibited 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 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**

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±0.4**	**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

Work partially supported by ONR, ARO, NGA, NSF, NIH, Simons Foundation, and gifts from AWS, Microsoft, and Google.

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.