PARETO OPTIMALITY IN NO-HARM FAIRNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Common fairness definitions in machine learning focus on balancing various notions of disparity and utility. In this work we study fairness in the context of risk disparity among sub-populations. We introduce the framework of *Pareto-optimal fairness*, where the goal of reducing risk disparity gaps is secondary only to the principle of not doing unnecessary harm, a concept that is especially applicable to high-stakes domains such as healthcare. We provide analysis and methodology to obtain maximally-fair no-harm classifiers on finite datasets. We argue that even in domains where fairness at cost is required, no-harm fairness can prove to be the optimal first step. This same methodology can also be applied to any unbalanced classification task, where we want to dynamically equalize the misclassification risks across outcomes without degrading overall performance any more than strictly necessary. We test the proposed methodology on real case-studies of predicting income, ICU patient mortality, classifying skin lesions from images, and assessing credit risk, demonstrating how the proposed framework compares favorably to other traditional approaches.

1 INTRODUCTION

Machine learning algorithms play an important role in decision making in society. When these algorithms are used to make high-impact decisions such as hiring, credit-lending, predicting mortality for intensive care unit patients, or classifying benign/malign skin lesions, it is paramount to guarantee that these decisions are 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. Much of the fairness in machine learning literature attempts to produce algorithms that satisfy Demographic Parity, which aims to make algorithm's predictions independent of the sensitive populations (Louizos et al. (2015); Zemel et al. (2013); Feldman et al. (2015)); or Equality of Odds or Equality of Opportunity, which aims to produce predictions that are independent of the sensitive attributes given the ground truth (Hardt et al. (2016); Woodworth et al. (2017)). These notions of fairness 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 any unnecessary harm to any subgroup (Ustun et al. (2019)). Even if the overall fairness goal is a potentially harmful, zero-gap classifier, pursuing first a Pareto-fair classifier and later applying other harmful methodologies ensures that all possible non-harmful trade-offs are covered before explicitly degrading performance, therefore minimal harm is introduced to the decision.

In this work we make use of the concept of Pareto optimality to measure and analyze discrimination (unfairness) in terms of the difference in predictive risks across sub-populations defined by our sensitive attributes, a fairness metric that has been explored in other recent works such as Calders & Verwer (2010); Dwork et al. (2012); Feldman et al. (2015); Chen et al. (2018); Ustun et al. (2019). We examine the subset of models from our hypothesis class that have the best trade-offs between sub-population risks, and select from this set the one with the smallest risk disparity gap. This is in direct contrast to common post-hoc correction methods like the ones proposed in Hardt et al. (2016); Woodworth et al. (2017), where noise is potentially added to the decisions of the best performing sub-population. While this latter type of approach diminishes the risk-disparity gap, it does so by degrading performance on advantaged groups, the previously disadvantaged groups do not directly benefit from this treatment. Since our proposed methodology does not require test-time access to sensitive attributes, and can be applied to any standard classification or regression task, it can also be used to reduce risk disparity between outcomes, acting as an adaptive risk equalization loss compatible with unbalanced classification scenarios.

Main Contributions. We formalize the notion of no-harm risk fairness using Pareto optimality (Mas-Colell et al. (1995)), a state of resource allocations from which it is impossible to reallocate without making one subgroup worse. We show that finding a Pareto-fair classifier is equivalent to finding a model in our hypothesis class that belongs to the Pareto front (the set of all Pareto optimal allocations) with respect to the sub-population risks with the smallest possible risk disparity. This general notion is already amenable to non-binary sensitive attributes. We analyze the fairness performance trade-offs that can be expected from different approaches with an illustrative example. We provide a concrete algorithm that promotes fair solutions belonging to the Pareto front; this algorithm can be applied to any standard classifier or regression task that can be trained using (Stochastic) Gradient Descent. We show that if the goal is to obtain a zero-gap classifier, first recovering the fairest Pareto optimal solution and then adding harmful post-hoc corrections ensures the lowest risk levels across all subgroups. We demonstrate how our methodology performs on synthetic and real tasks such as inferring income status in the Adult dataset Dua & Graff (2017a) irrespective of their ethnicity or gender, 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).

2 Related Work

There is an extensive body of work on fairness in machine learning. Following Friedler et al. (2019), we compare our methodology against the works of Feldman et al. (2015); Kamishima et al. (2012); Zafar et al. (2017). Our method shares conceptual similarities with Zafar et al. (2017); Woodworth et al. (2017); Agarwal et al. (2018), with differences on how we define our fairness objective and adapt it to work with standard neural networks. Although optimality is often discussed in the fairness literature, it is usually in the context of utility-discrimination tradeoffs. To the best of our knowledge, this is the first work to discuss optimality with respect to subgroup risks on a unified classifier, a distinction that disallows extreme performance degradation in the pursuit of fairness.

The work presented in Hashimoto et al. (2018) discusses decoupled classifiers 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 because it allows transfer learning between diverse sub-populations. We do not need access to the sensitive attribute during test time, but in cases where this is possible, we instead choose to incorporate it as part of our observation features.

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 subclass. Their methodology attempts to bridge the disparity gap by collecting additional samples from high-risk sub-populations. Here we modify our classifier loss to bridge the disparity gap without inducing unnecessary harm, which could prove to be synergistic with their methodology.

3 Problem Statement

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, etc.), $y_i \in \mathcal{Y}$ 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.

Let $h \in \mathcal{H}$ be a classifier from a compact hypothesis class $\mathcal{H}$ trained to infer y from x, $h : \mathcal{X} \to \mathcal{Y}$; and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. We define the class-specific risk of classifier h on subgroup a as $R_a(h) = \mathbb{E}_{X,Y|A=a}[\ell(Y, h(X))]$. The risk discrimination gap between two subgroups $a, a' \in \mathcal{A}$ is measured as $\Gamma_{a,a'}(h) = |R_a(h) - R_{a'}(h)|$, and we define the pairwise discrimination gap as $\vec{\Gamma}_{\mathcal{A}}(h) = \{\Gamma_{a,a'}(h)\}_{a,a' \in \mathcal{A}}$. Our goal is to obtain a classifier $h \in \mathcal{H}$ that minimizes this gap without causing unnecessary harm to any particular group in $\mathcal{A}$. To formalize this notion, we define:

Definition 3.1. Dominant classifier: Classifier h' is said to dominate h, noted as $h' \succ h$, if $R_a(h) \geq R_a(h'), \forall a \in \mathcal{A}$ and $\exists\, a' \in \mathcal{A} : R_{a'}(h) > R_{a'}(h')$ (i.e., strict inequality on at least one group risk).
Definition 3.2. Pareto front classifier: A classifier h belongs to the Pareto front $\mathcal{P}(\mathcal{H}, \mathcal{A})$ if it is not dominated by any $h' \in \mathcal{H}$. We formally define it as $\mathcal{P}(\mathcal{H}, \mathcal{A}) = \{h \in \mathcal{H} : \nexists h' \in \mathcal{H}, \nexists a' \in \mathcal{A} : [R_{a'}(h') < R_{a'}(h)] \wedge ([R_a(h') \leq R_a(h)], \forall a \in \mathcal{A})\}$. This means that there is no other classifier in $\mathcal{H}$ that is at least as good in all risks and strictly better in at least one of them. It is the set of classifiers such that improving one group's risk comes at the cost of increasing another's.

The Pareto front defines the best achievable trade-offs between population risks $R_a(h)$. This definition 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 class-specific risks and strictly better in one of them. As shown in Chen et al. (2018); Domingos (2000), the risk can be decomposed in bias, variance and noise for some loss functions, where the noise is the smallest achievable risk for infinitely large datasets (Bayes-optimal risk). If the noise differs between sensitive groups, zero-discrimination (perfect fairness) can only be achieved by introducing bias or variance.

The following, Lemma 3.1, shows that applying a mechanism for reaching equality of risk on a dominated classifier $h \in \mathcal{H}$ leads to equal or worse risks than applying it to a Pareto optimal classifier $h_p \in \mathcal{P}(\mathcal{H}, \mathcal{A})$ that dominates h. This motivates searching for Pareto-efficient classifiers, potentially as an intermediate step, even when the ultimate goal is a degraded, zero-disparity classifier.
Lemma 3.1. *If $h \notin \mathcal{P}(\mathcal{H}, \mathcal{A}) \rightarrow \exists\, h_p \in \mathcal{P}(\mathcal{H}, \mathcal{A}) : h_p \succ h \wedge R_a(h_p^{ER}) \leq R_a(h^{ER})\, \forall a$, with h^{ER} an equal-risk classifier : $R_a(h^{ER}) = \max\limits_{a' \in \mathcal{A}} R_{a'}(h), \forall a$ and $h_p^{ER} : R_a(h_p^{ER}) = \max\limits_{a' \in \mathcal{A}} R_{a'}(h_p)$.*

Literature on fairness has focused on putting constraints on the norm of discrimination gaps similar to $||\vec{\Gamma}_{\mathcal{A}}(h)||_\infty$ (Zafar et al. (2017; 2015); Creager et al. (2019); Woodworth et al. (2017)). We follow a similar criteria in Definition 3.3 and define the optimal Pareto-fair classifier as the Pareto-optimal one that minimizes this norm. Note that one could alternatively choose to find the Pareto classifier that minimizes the maximum subgroup risk.
Definition 3.3. Optimal Pareto-fair classifier and Pareto-fair point: A classifier h^* is an optimal Pareto-fair classifier if it minimizes the discrimination gap among all Pareto front classifiers, $h^* = \underset{h \in \mathcal{P}(\mathcal{H}, \mathcal{A})}{\arg\min}\ ||\vec{\Gamma}_{\mathcal{A}}(h)||_\infty$. The Pareto-fair point is defined as $\vec{R}^* = \{R_a(h^*)\}_{a \in \mathcal{A}}$.

Figure 1 shows a scenario with binary sensitive attributes a and binary output variable y where none of the classifiers in the Pareto front intersect the equality of risk line. Here the level of noise between subgroups differs, and the Pareto-fair point $\vec{R}^*$ would not be achieved by either a Naive classifier (minimizes expected global risk), or a classifier where subgroups are re-sampled to appear with equal probability (rebalanced Naive classifier). Note that the amount of performance degradation required to enforce perfect fairness starting from the Naive classifier is significantly higher than when starting from the Pareto-fair point.

Our objective is to find the optimal no-harm classifier h^* as in Definition 3.3. In order to guarantee that the Pareto front $\mathcal{P}(\mathcal{H}, \mathcal{A})$ is non-empty, we assume that for any functions $\phi(\{R_a(h)\}) : \mathbb{R}^{|\mathcal{A}|} \rightarrow \mathbb{R}$ that are monotonically increasing in the risks, as is the case of the Naive loss $\phi^l_{\text{naive}}(\{R_a(h)\}) = \sum\limits_{a \in \mathcal{A}} P_a R_a(h)$, the classifier $h' = \underset{h \in \mathcal{H}}{\arg\min}\, \phi(\{R_a(h)\})$ is also in the hypothesis class (not an infimum). To simplify our analysis further, we also assume that the Pareto front is compact, this ensures that h^* is a minimum and not an infimum. The following lemma formalizes these statements and shows that we can recover classifiers in the Pareto front with different levels of discrimination by minimizing $\phi(\{R_a(h)\})$ subject to a discrimination constraint.
Lemma 3.2. *Given that $\exists \underset{h \in \mathcal{H}}{\arg\min}\, \phi(\{R_a(h)\}_{a \in \mathcal{A}})$ for any $\phi(\{R_a(h)\}_{a \in \mathcal{A}}) : \mathbb{R}^{|\mathcal{A}|} \rightarrow \mathbb{R}$ monotonically increasing with $R_a(h), \forall a$ and $\epsilon^* = \underset{h \in \mathcal{P}(\mathcal{H}, \mathcal{A})}{\min} ||\vec{\Gamma}_{\mathcal{A}}(h)||_\infty$ we have the following:*

$$\begin{aligned} \bar{h} &= \underset{h \in \mathcal{H}}{\arg\min}\, \phi(\{R_a(h)\}_{a \in \mathcal{A}}) \textit{ belongs to } \mathcal{P}(\mathcal{H}, \mathcal{A}); \\ h' &= \underset{h \in \mathcal{H}}{\arg\min}\, \phi(\{R_a(h)\}_{a \in \mathcal{A}})\ s.t. ||\vec{\Gamma}_{\mathcal{A}}(h)||_\infty \leq \epsilon' \textit{ belongs to } \mathcal{P}(\mathcal{H}, \mathcal{A}), \forall \epsilon' \geq \epsilon^*; \\ \hat{h} &= \underset{h \in \mathcal{H}}{\arg\min}\, \phi(\{R_a(h)\}_{a \in \mathcal{A}})\ s.t. ||\vec{\Gamma}_{\mathcal{A}}(h)||_\infty \leq \epsilon^* \textit{ is a Pareto-fair classifier}. \end{aligned} \tag{1}$$

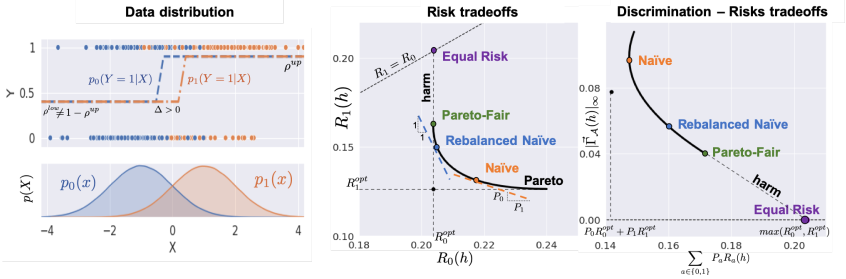

Figure 1: Example of binary output variable $y = \{0, 1\}$ and two sensitive groups $a = \{0, 1\}$, with $P_{a=1} > P_{a=0}$. Bottom left figure shows conditional distributions of observation variable $p_a(x)$. Top left shows output variable distribution as a function of both observation and sensitive variable $p(y = 1|x, a)$ which for simplicity are piece-wise constants with two levels ρ^{low} and ρ^{high}, Δ represents the gap between the level transitions for each group. Middle figure shows the Pareto frontier between (subgroup) risks for this problem; since noise levels are not the same across subgroups, perfect fairness is not attainable. Equality of Risk requires pure degradation of service for group $a = 1$. Both Naive and Rebalanced Naive classifiers do not attain the best possible no-harm classifier in this case. The proposed Pareto-fair point is shown in green. The utopic point (R_0^{opt},R_1^{opt}) can only be reached in the infinite sample regime where the classifier also has access to the sensitive attribute ($h(X, A)$). Rightmost figure shows the trade-offs attainable between discrimination and mean risks, an equivalent problem to the risk trade-off figure.

Proofs for all lemmas are given in the supplementary material, Section A.1. In Section 4 we provide a concrete algorithm to approximately recover the optimal no-harm classifier h^* by adapting a standard Stochastic Gradient Descent approach.

4 Optimization Methods

Recall that we wish to recover the Pareto-fair classifier h^* within our hypothesis class. From Lemma 3.2 we can recover this classifier by solving $\hat{h} = \arg\min_{h \in \mathcal{H}} \phi(\{R_a(h)\}_{a \in \mathcal{A}})\ s.t. ||\vec{\Gamma}_{\mathcal{A}}(h)||_\infty \leq \epsilon^*$, for any monotonic function ϕ. Equivalently, we can also solve $\hat{h} = \arg\min_{h \in \mathcal{H}} \phi(\{R_a(h)\}_{a \in \mathcal{A}})\ s.t. : ||R_a(h) - \min_{a' \in \mathcal{A}} R_{a'}(h)|| < \epsilon^*, \forall a \in \mathcal{A}$. One approach to solving this numerically is to use the squared penalty loss Nocedal & Wright (2006),

$$\hat{\phi} = \sum_{a \in \mathcal{A}} R_a(h) + \mu_a \max(R_a(h) - \min_{a' \in \mathcal{A}} R_{a'}(h) - \epsilon^*, 0)^2, \tag{2}$$

where we used $\sum_{a \in \mathcal{A}} R_a(h)$ as an example of the monotonically decreasing function. The two main challenges of this approach are that ϵ^* is unknown and that the weights μ_a need to be dynamically adjusted, which is tackled as follows.

From Lemma 3.2 we have that for any monotonically decreasing loss $\phi(\{R_a(h)\})$, $h = \arg\min_{h \in \mathcal{H}} \phi(\{R_a(h)\})$ already belongs to the Pareto front of the classifier. Since we wish to avoid searching for ϵ^* and we want to ensure our classifier is in the Pareto boundary, we define the loss

$$\phi(h; \vec{\mu}, h^-) \quad = \sum_{a \in \mathcal{A}} R_a(h) + \mu_a \max(R_a(h) - \min_{a'} R_{a'}(h^-), 0)^2. \tag{3}$$

Here we make use of a target classifier h^-, which is a delayed copy of our current classifier h. Having a target classifier h^- implies that no gradients from term $\min_{a'} R_{a'}(h^-)$ affect the current network parameters, making the loss function $\phi(h; \vec{\mu}, h^-)$ monotonically decreasing for all $\vec{\mu} \succ 0$. Target networks in loss functions are extensively used in the domain of Reinforcement Learning to reduce overestimation Van Hasselt et al. (2016), we adapt this idea to enforce monotonicity. The penalty terms $\vec{\mu}$ are dynamically adjusted until no further reduction of the disparities $R_a(h) - \min_{a'} R_{a'}(h^-)$ is seen on validation data.

The proposed implementation of the Pareto-fair framework is formalized in Algorithm 1 (shown in supplementary material, Section A.2) where we specify how to update the penalty coefficients

μ_a. We regularly check that reductions in the fairness gap obtained on the training set generalize by evaluating this gap on the validation set; we additionally check if the trade-offs are in the non-dominated solution set (i.e., we have not observed a universally better classifier during training). Algorithm 2 (shown in supplementary material, Section A.2) summarizes how we perform stochastic gradient descent steps with early stopping in-between μ_a update steps.

To conclude this section let us stress that the proposed framework is independent of the desired algorithm class $\mathcal{H}$ and risk R; these are kept from the original application. The Pareto-fair classifier uses the same inputs as the Naive classifier, with parameters that have been optimized towards no-harm fairness. Code will be made available.

5 Experiments and Results

We applied the methodology described in Section 4 to learn a classifier (from hypothesis class $\mathcal{H}$) in the group-risk Pareto front with the smallest group-risk disparity (Pareto-fair classifier). We first validate our methodology on synthetic data with known optimal Pareto-fair classifiers. Observations are drawn from a Gaussian mixture model where each sensitive attribute is encoded by a corresponding Gaussian mode, and target attributes are binary. We demonstrate our methodology on standard publicly available fairness datasets, and show how the risk disparity gaps are subsequently reduced. Where applicable, we compare our results against the methodologies proposed in Zafar et al. (2017); Kamishima et al. (2012); Hardt et al. (2016); Feldman et al. (2015).

5.1 Synthetic Data

We tested our approach on synthetic data where the observations are drawn from conditional Gaussian distributions $X|A = a \sim N(\mu_A, 1)$, the target variable y is a conditional Bernoulli variable with distribution $Y|X = x, A = a \sim Ber\big(f_a(x)\big)$ with $f_a(x) = \rho_a^{low}\mathbb{1}[x \leq c_a] + \rho_a^{high}\mathbb{1}[x > c_a]\big)$, and $|A| = 3$. We used mean squared error (MSE) as our loss function. Under these conditions, the Bayes-optimal classifier for each subgroup is $h(x) = f_a(x)$. The sub-group risk function for any classifer h can be computed as $R_a(h) = \mathbb{E}_{X|A=a}[Y - h(X))^2] = \mathbb{E}_{X|A=a}[f_a(X)(1 - h(X))^2 + (1 - f_a(X))h(X)^2]$. Network details are given in the supplementary material, Section A.3. Figure 2 shows the analytically-derived Pareto front (see Section A.4 in supplementary material), as well as the trade-off obtained by the proposed algorithm and a balanced Naive classifier.

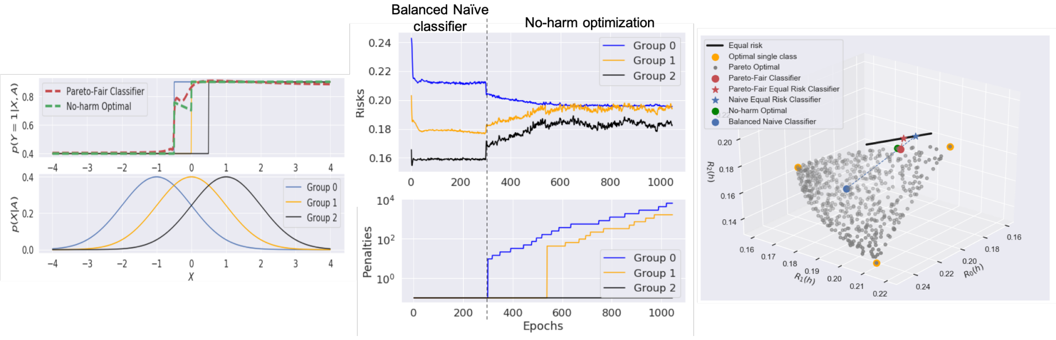

Figure 2: Synthetic data experiment. Bottom left figure shows the conditional distribution of the observation variables for each of the three subgroups, while upper left shows the distribution of the target variable conditioned on both the observation and sensitive attribute; it additionally shows the theoretical and empirical Pareto-fair classifier. Center figures shows the empirical validation risks as a function of epochs during the optimization procedure, with a sharp transition visible when we begin following Algorithm 1 (top) and the values of the penalties at every epoch (bottom). Rightmost figure shows the risk distribution of all Pareto-optimal classifiers, as well as the risk attained by a rebalanced Naive classifier, and the classifier recovered using our method (close to the optimum); the equality of risk line is shown for reference.

Our approach works as expected, we can see that the Pareto-fair optimization procedure is able to correctly trade risks from the two advantaged subgroups to reduce the risk of the worst performing subgroup. It is also able to obtain a classifier that is close to the theoretical optimal by learning from finite samples, and outperforms the balanced Naive classifier.

5.2 Real Datasets

We evaluate group-risk disparity gaps on mortality prediction, skin lesion classification, income prediction, and credit lending for fairness-unaware (Naive) classifiers. We demonstrate how the proposed Pareto-fair approach can reduce those risk disparity gaps without unnecessary harm using the proposed optimization methodology.

Similarly to Chen et al. (2018); Zafar et al. (2017); Hardt et al. (2016); Woodworth et al. (2017), we omit the sensitive attribute $a \in \mathcal{A}$ from our observation features. Our method trains a single classifier for the entire dataset to avoid needing test-time access to sensitive attributes whenever possible. All classifiers shown in this section are implemented using neural networks and/or logistic regression, we used either cross-entropy or categorical MSE as our training loss depending on the dataset, and a 60/20/20 train-validation-test split. For details on the architecture used on each dataset, refer to the supplementary material, Section A.3.

We compare our methodology against a Naive classifer, where no special consideration is given to fairness, a rebalanced Naive classifier, where we oversample underrepresented sub-populations, and against the post-processing framework presented in Hardt et al. (2016). We also apply this latter framework to the no-harm classifier recovered by our method, which can be seen as an extension and experimental validation of the ideas advanced in Woodworth et al. (2017). Additional comparisons are included when possible.

5.2.1 Predicting Mortality in Intensive Care Patients

Medical decisions in general and mortality prediction in particular are examples where notions of fairness among sub-populations are of paramount importance, and where ethical considerations make no-harm fairness a very attractive paradigm. To that end, 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 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 where 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 study fairness with respect to age (under/over 55 years old), self-reported ethnicity, and outcome. We chose to include the outcome (Alive/Deceased) variable as a sensitive sub-population criteria both to demonstrate a case where sensitive attributes would not be available at test-time, and also because in our experiments patients who ultimately passed away on ICU were under-served by a Naive classifier, with predictive risks significantly higher than surviving patients. Table 1 shows empirical risks and accuracy of all tested methodologies.

5.2.2 Skin Lesion Classification

The HAM10000 dataset Tschandl et al. (2018) collects over $10,000$ dermatoscopic images of skin lesions over a diverse population. Lesions are classified according to diagnostic categories. Diagnosis is confirmed by expert consensus, pathology, and follow-up appointments or confocal microscopy. Gender and age attributes of participants were also recorded.

We trained a DenseNet121 network Huang et al. (2017) to classify skin lesions from dermatoscopic images. We found that a Naive classifier exhibited almost no measurable discrimination based on age or race on this dataset. We instead chose to use the diagnosis class as both the target and sensitive variable, casting balanced risk minimization as a particular use-case for Pareto fairness. This is possible since our methodology does not require test-time access to sensitive labels. It was not possible to show comparisons against Hardt et al. (2016) since the sensitive attribute is perfectly predictive of the outcome. Table 2 shows empirical risks and accuracies for all tested methodologies.

5.2.3 Income Prediction and Credit Risk Assesment

We tested the proposed method on the Adult UCI dataset Dua & Graff (2017a) and on the German Credit dataset Dua & Graff (2017b). In the Adult UCI dataset the goal is to predict a person's income, which can be an important factor on meaningful decisions such as credit lending. This dataset contains 105 binarized observations representing education status, age, ethnicity, gender, and marital status and a target variable indicating income status (binary attribute representing over

Out/Age/Race	Naive	Rebalanced Naive	Naive+Zafar	Ours	Rebalanced Naive+Hardt	Ours+Hardt
A/A/NW	0.07 / 96.0%	0.06 / 94.9%	0.01 / 100.0%	0.20 / 85.3%	68.9%	68.2%
A/A/W	0.06 / 96.2%	0.06 / 95.1%	0.01 / 100.0%	0.21 / 82.4%	70.4%	70.7%
A/S/NW	0.10 / 94.8%	0.13 / 87.9%	0.01 / 100.0%	0.32 / 70.3%	69.7%	70.6%
A/S/W	0.09 / 94.1%	0.14 / 87.7%	0.01 / 99.8%	0.31 / 72.5%	69.8%	70.2%
D/A/NW	0.67 / 44.0%	0.74 / 52.0%	2.27 / 28.0%	0.34 / 80.0%	65.1%	73.2%
D/A/W	0.48 / 50.0%	0.58 / 54.5%	0.60 / 71.2%	0.15 / 89.4%	48.9%	79.7%
D/S/NW	0.58 / 35.7%	0.61 / 42.6%	0.33 / 89.6%	0.23 / 75.7%	51.3%	74.6%
D/S/W	0.57 / 37.5%	0.60 / 52.9%	0.05 / 98.8%	0.23 / 76.4%	53.8%	73.9%
Mean	0.33 / 68.5%	0.37 / 71.0%	0.41 / 85.9%	0.25 / 79.0%	62.2%	72.6%
Discrimination	0.61 / 60.5%	0.68 / 52.5%	2.27 / 72.0%	**0.18 / 19.1%**	21.5%	**11.5%**

Table 1: Group-specific cross-entropy risks and accuracies for Naive, Rebalanced Naive, Zafar applied to the features learned in the Naive classifier (Naive+Zafar), and our proposed Pareto-Fair classifiers (Ours). We considered the combination of Outcome (Alive or Deceased, A/D), Age (Adult or Senior, A/S) and Race (White or Non-White, W/NW) to be our sensitive attributes, and Outcome as our target attribute. No classifier had test-time access to the sensitive attributes. Accuracies for the post-processing Equality of Odds method proposed in Hardt et al. (2016) applied on the Rebalanced Naive (Rebalanced Naive+Hardt) and Pareto-Fair (Ours+Hardt) classifiers are also shown (equalized across Age and Ethnicity). The proposed Pareto-Fair classifier exhibited the lowest balanced mean risk and risk disparity amongst all trained classifiers. Although Naive+Zafar exhibits the highest mean accuracy, it does so at the cost of severly underserving non-white adults who ultimately passed away, as a result, their discrimination values are relatively high. Only the Ours+Hardt classifier exhibits lower accuracy disparity, but the reduction in accuracy disparity is similar to the reduction in mean accuracy. When comparing Ours+Hardt and Rebalanced Naive+Hardt classifiers, we note that not only the resulting accuracy discrimination is lower on the Ours+Hardt classsifier, the mean accuracy is also significantly higher, supporting the argument of first applying Pareto-fairness before doing harmful post-processing.

Lesion Type	Naive	Rebalaced Naive	Ours
akiec	0.110 / 48%	0.118 / 56%	0.089 / 63%
bcc	0.096 / 60%	0.059 / 73%	0.065 / 63%
bkl	0.110 / 51%	0.105 / 58%	0.092 / 52%
df	0.142 / 50%	0.143 / 33%	0.053 / 83%
mel	0.011 / 95%	0.009 / 96%	0.024 / 89%
nv	0.043 / 86%	0.057 / 71%	0.044 / 79%
vasc	0.156 / 37%	0.125 / 50%	0.069 / 65%
Mean	0.095 / 61%	0.088 / 63%	0.062 / 71%
Discrimination	0.145 / 58%	0.134 / 63%	**0.068 / 37%**

Table 2: Group-specific mean square error risks and accuracies for Naive Rebalanced Naive and proposed Pareto-Fair (Ours) classifiers. We took Lesion Type to be both our target and sensitive variable; a particular application of Pareto-fairness to adaptive risk-equalization. Lesion types are classified as Actinic keratoses and intraepithelial carcinoma (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv) and vascular lesions (vasc). The Pareto-fair classifier behaves as expected, exhibiting the best balanced mean risk and balanced accuracy, as well as the lowest risk and accuracy discriminations among the three tested methodologies. Note that comparisons against the other baselines is not possible since the target label is not binary.

or under $\$50,000$); data is collected on $32,561$ adults. In the German Credit dataset the goal is predicting credit risk, this dataset contains $1,000$ entries with 20 observation attributes; sensitive attribute gender is not directly included in the data, but can be derived from the given information.

We strive to ensure no-harm fairness across gender and ethnicity. To compare ourselves against state of the art methods Zafar et al. (2017); Feldman et al. (2015); Kamishima et al. (2012) we binarize the sensitive attributes into White-Male and Other when dealing with ethnicity and gender simultaneously, or Male and Female when dealing with gender, and use the unified testbed provided in Friedler et al. (2019). We limit our hypothesis class $\mathcal{H}$ to linear logistic regression to compare evenly against these standard baselines. Results on the Adult dataset when repairing for ethnicity and gender are shown in Table 3a. Comparable results when repairing for only the gender attribute in both the Adult and German Credit datasets are shown in tables 3b and 3c respectively. All values are reported on test data using 5-fold cross-validation. Non-Hardt methods show similar accuracy performances on these particular examples, with our method showing consistently smaller risk disparity gaps. Note that no other method has group risks that dominate our Pareto-fair classifier, the converse is not true, all other classifiers have risks that are strictly dominated by our classifier, suggesting their results are not Pareto efficient. As expected, Hardt is the method with the smallest accuracy disparity, but this comes at a steep accuracy cost. In the Adult dataset, where there is sufficient data to train a neural network, we observe that it slightly outperforms linear logistic regression as expected. Lower cross-entropy risks do not necessarily translate to highest accuracies, this was especially true on the relatively small German Credit dataset.

Sensitive Attribute	Hardt	Kamishima	Feldman	Zafar	Ours-LR	Ours-NN	Ours+Hardt-NN
White Male	78.2%	91.0%	90.8% / 3.17	90.5% / 0.24	90.5% / 0.11	90.8% / 0.12	80.8%
Other	74.6%	80.5%	80.4% / 2.47	80.1% / 0.42	80.7% / 0.21	81.0% / 0.20	78.3%
Mean	76.4%	85.7%	85.6% / 2.82	85.3% / 0.33	85.6% / 0.16	85.9% / 0.16	79.6%
Discrimination	3.5%	10.5%	10.4% / 0.70	10.3% / 0.18	9.8% / 0.10	9.8% / 0.07	2.5%

(a) Method comparison in Adult dataset with sensitive groups White Male and Other.

Sensitive Attribute	Hardt	Kamishima	Feldman	Zafar	Ours-LR	Ours-NN	Ours+Hardt-NN
Male	82.9%	92.6%	92.3% / 3.19	92.2% / 0.20	91.9% / 0.10	91.9% / 0.11	81.8%
Female	79.6%	80.9%	80.7% / 2.44	80.8% / 0.41	81.0% / 0.20	81.6% / 0.20	78.8%
Mean	81.2%	86.7%	86.5% / 2.81	86.5% / 0.30	86.4% / 0.15	86.7% / 0.15	80.3%
Discrimination	3.3%	11.7%	11.6% / 0.74	11.4% / 0.20	11.9% / 0.10	10.3% / 0.09	3.0%

(b) Method comparison in Adult dataset with sensitive groups Male and Female.

Sensitive Attribute	Hardt	Kamishima	Feldman	Zafar	Ours-LR
Male	50.7%	63.7%	73.2% / 0.57	72.5% / 0.57	65.1% / 0.36
Female	50.4%	70.5%	71.1% / 0.56	71.6% / 0.63	68.8% / 0.38
Mean	50.5%	67.1%	72.1% / 0.57	72.1% / 0.60	67.0% / 0.37
Discrimination	0.4%	6.8%	2.2% / 0.01	0.9% / 0.06	3.7% / 0.02

(c) Method comparison in German Credit dataset with sensitive groups Male and Female.

Table 3: Accuracy and cross-entropy results on Adult dataset with sensitive groups White Male and Other (a), Male and Female (b), and on German Credit dataset with sensitive attributes Male and Female (c); no classifier had test-time access to the sensitive attributes. We compare against the methods proposed in Hardt et al. (2016) (Hardt),Koski (1985) (Kamishima), Feldman et al. (2015) (Feldman), Zafar et al. (2017) (Zafar); all of which use linear logistic regression as their hypothesis class. We show accuracy results on our method using linear logistic regresion (OursLR), a fully-connected neural network (Ours-NN), and postprocessing on the neural network classifier (Ours+Hardt-NN). Fairness parameters for Kamishima, Feldman and Zafar were adjusted by sweeping, results report performance of smaller risk disparity classifier.

6 Discussion

There exists a rich literature of fairness in machine learning in general, and risk-based fairness in particular. Here we explore a relatively untapped sub-problem where the goal is to reduce risk disparity gaps in the most ethical way possible (i.e., minimizing unnecessary harm). Unlike other works in the area, our problem investigates on how to reduce this disparity gap without collecting additional data samples, using the entirety of the available training data to produce an algorithm that is maximally fair with respect to sub-populations, and that does not necessarily require test-time access to sensitive attributes.

We provide a concrete algorithmic adaptation to any standard classification or regression loss to bridge this disparity gap at no unnecessary harm, and demonstrate its performance on several real-world case studies. Even for applications where the need for strict fairness outweighs the need for no-harm classifiers, this methodology can be applied before any post-hoc corrections to ensure that the risk disparity gap is closed in the most risk-efficient way for all involved sub-populations. The proposed algorithm does not sweep through different disparity constraint values, as previously done in related works, making it a simpler alternative.

As an avenue of future research, it could be of interest to analyze if we can 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 or populations. We strongly believe that no-harm notions of fairness are of great interest for several applications, especially so on domains such as healthcare and lending, where decisions have highly impactful.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.

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

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

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

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. *arXiv preprint arXiv:1906.02589*, 2019.

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

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

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

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226. ACM, 2012.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 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.

Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 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.

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

Moritz Hardt, Eric Price, Nathan Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

Tatsunori B Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. *arXiv preprint arXiv:1806.08010*, 2018.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

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

Juhani Koski. Defectiveness of weighting method in multicriterion optimization of structures. *Communications in applied numerical methods*, 1(6):333–337, 1985.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pp. 6402–6413, 2017.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

Andreu Mas-Colell, Michael Dennis Whinston, Jerry R Green, et al. *Microeconomic theory*, volume 1. Oxford university press New York, 1995.

Kaisa Miettinen. *Nonlinear multiobjective optimization*, volume 12. Springer Science & Business Media, 2012.

Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

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

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

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Thirtieth AAAI conference on artificial intelligence*, 2016.

Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*, 2017.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.

Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pp. 229–239, 2017.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.

A Appendix

A.1 Proofs

Here we restate the Lemmas shown in Section 3 along with a sketch of the proofs

Lemma 3.1 If $h \notin \mathcal{P}(\mathcal{H},\mathcal{A}) \rightarrow \exists\, h_p \succ h \in \mathcal{P}(\mathcal{H},\mathcal{A}) : R_a(h_p^{ER}) \leq R_a(h^{ER}), \forall a$, where h^{ER} is a equality of risks classifier in $\mathcal{H}$ such that $R_a(h^{ER}) = \max_{a' \in \mathcal{A}} R_{a'}(h), \forall a$ and $h_p^{ER} : R_a(h_p^{ER}) = \max_{a' \in \mathcal{A}} R_{a'}(h_p), \forall a$

Proof. If h_p dominates $h \rightarrow R_a(h^{ER}) = \max_{a' \in \mathcal{A}} R_{a'}(h_p) \leq \max_{a' \in \mathcal{A}} R_{a'}(h) = R_a(h^{ER})$ □

Lemma 3.2 Given that $\exists \arg\min_{h \in \mathcal{H}} \phi(\{R_a(h)\}_{a \in \mathcal{A}})$ for any $\phi(\{R_a(h)\}_{a \in \mathcal{A}}) : \mathbb{R}^{|\mathcal{A}|} \rightarrow \mathbb{R}$ monotonically increasing with $R_a(h), \forall a$ and $\epsilon^* = \min_{h \in \mathcal{P}(\mathcal{H},\mathcal{A})} ||\vec{\Gamma}_{\mathcal{A}}(h)||_\infty$ we have the following:

$$\begin{aligned} \bar{h} &= \arg\min_{h \in \mathcal{H}} \phi(\{R_a(h)\}_{a \in \mathcal{A}}) \text{ belongs to } \mathcal{P}(\mathcal{H},\mathcal{A}); \\ h' &= \arg\min_{h \in \mathcal{H}} \phi(\{R_a(h)\}_{a \in \mathcal{A}}) \ s.t. ||\vec{\Gamma}_{\mathcal{A}}(h)||_\infty \leq \epsilon' \text{ belongs to } \mathcal{P}(\mathcal{H},\mathcal{A}), \forall \epsilon' \geq \epsilon^*; \\ \hat{h} &= \arg\min_{h \in \mathcal{H}} \phi(\{R_a(h)\}_{a \in \mathcal{A}}) \ s.t. ||\vec{\Gamma}_{\mathcal{A}}(h)||_\infty \leq \epsilon^* \text{ is a Pareto-fair classifier.} \end{aligned} \tag{4}$$

Proof. By definition, $\epsilon^* = \min_{h \in \mathcal{P}(\mathcal{H},\mathcal{A})} ||\vec{\Gamma}_{\mathcal{A}}(h)||_\infty$, therefore, $\forall \epsilon' \geq \epsilon^*$ we know that the set of all feasible classifiers that belong to both the Pareto front $\mathcal{P}(\mathcal{H},\mathcal{A})$ and satisfy $||\vec{\Gamma}_{\mathcal{A}}(h)||_\infty \leq \epsilon'$ is non-empty, call this set $\mathcal{P}(\mathcal{H},\mathcal{A},\epsilon')$. By contradiction, suppose $h' \notin \mathcal{P}(\mathcal{H},\mathcal{A},\epsilon')$, therefore, there exists a classifier $\tilde{h} \in \mathcal{P}(\mathcal{H},\mathcal{A},\epsilon')$, and a subgroup $\tilde{a}$ such that $R_a(h') \geq R_a(\tilde{h}), \forall a \in \mathcal{A}$ and $R_{\tilde{a}}(h') > R_{\tilde{a}}(\tilde{h})$. Since $\phi(\{R_a(h)\}_{a \in \mathcal{A}})$ is monotonically decreasing in $R_a(h)$, then $\phi(\{R_a(h')\}) > \phi(\{R_a(\tilde{h})\})$, which is absurd, proving that $h' \in \mathcal{P}(\mathcal{H},\mathcal{A})$

The statements for $\hat{h}$ and $\bar{h}$ follow from taking $\epsilon' = \epsilon^*, \infty$ respectively.

□

A.2 Algorithmic details

Here we provide the core Pareto-Fair algorithm (Algorithm 1), updating and optimizing the loss function proposed in Eq 3, a detailed view on how we optimize our adaptive loss in-between penalty update (μ_a) steps is shown in Algorithm 2.

Algorithm 1: ParetoFairOptimization

```
[h] Given: $h_\theta, L, \mathcal{D}^{\text{Tr}}, \mathcal{D}^{\text{Val}}, n_\mu, n_p, n_{max}, \gamma > 1, \text{lr}, \text{B}$
$\vec{\mu} \leftarrow \vec{0},\ \mu_{\text{count}} \leftarrow 0,\ e_{\text{count}} \leftarrow 0,\ \epsilon^* \leftarrow \infty,\ h^* \leftarrow h_\theta$
while $e_{count} \leq n_{max}$ and $\mu_{count} \leq n_\mu$ do
    $\mu_{\text{count}} \leftarrow \mu_{\text{count}} + 1,\ e_{\text{count}} \leftarrow e_{\text{count}} + 1$
    $h_\theta, R^{\text{Val}}_{-} \leftarrow \text{AdaptiveOptimize}(h_\theta, L, \vec{\mu}, \mathcal{D}^{\text{Tr}}, \mathcal{D}^{\text{Val}}, n_p, \text{lr}, \text{B})$ // Optimize loss with fixed $\vec{\mu}$

                    // Check that solution is no-harm and reduces fairness gap
    if $||\vec{\Gamma}_{\mathcal{A}}(h)||_\infty \leq \epsilon^*$ and $\vec{\bar{R}}^{val}(h)$ is not dominated by previous validation risks then
        $\mu_{\text{count}} \leftarrow 0,\ h^* \leftarrow h_\theta,\ \epsilon^* \leftarrow ||\vec{\Gamma}_{\mathcal{A}}(h)||_\infty$
    end
    $a' = \arg\max_a \bar{R}^{\text{Val}}_a(h)$                                  // Update $\mu$ values
    if $\mu_{a'} = 0$ then $\mu_{a'} \leftarrow \frac{\bar{R}^{\text{val}}_{a'}(h)}{2\epsilon^*}$ else $\mu_{a'} \leftarrow \gamma\mu_{a'}$
end
    // Exit loop due to excessive iterations or no improvement in fairness
Return: $h^*$
```

Algorithm 2: AdaptiveOptimize

```
Given: $h_\theta, L, \vec{\mu}, \mathcal{D}^{\text{Tr}}, \mathcal{D}^{\text{Val}}, n_p, \text{lr}, \text{B}$
$p \leftarrow 0, \quad \phi^* \leftarrow \infty \quad h^* \leftarrow h_\theta$
$R^{\text{Tr}}_{-} \leftarrow \min_a \bar{R}^{\text{Tr}}_a(h)$
while $p \leq n_p$                                              // Making loss progress
 do
    for $\{(x_i, y_i, a_i\}_{i=1}^{B} \in \mathcal{D}^{Tr}$                // Run one epoch of SGD on training set
     do
        $I_a = \{i \in \{1, \ldots, B\} \wedge a_i = a\}, \quad \bar{R}^{\text{B}}_a(h_\theta) = \frac{1}{|I_a|} \sum_{i \in I_a} L(h_\theta(x_i), y_i)$  // Empirical risks
        $\phi^B(h_\theta) = \sum_{a \in \mathcal{A}} \bar{R}^{\text{B}}_a(h_\theta) + \mu_a \max(\bar{R}^{\text{B}}_a(h_\theta) - R^{\text{Tr}}_{-}, 0)^2$
        $\theta \leftarrow \theta - lr\nabla_\theta \phi^B(h_\theta)$                              // Gradient step
    end
    if $\phi^{Val}(h_\theta) < \phi^*$        // Evaluate improvement on Val and update target risks
     then
        $h^* \leftarrow h_\theta; \quad \phi^* \leftarrow \phi^{\text{Val}}(h_\theta); \quad p \leftarrow 0$
    else
        $p \leftarrow p + 1;$
    end
    $R^{\text{Tr}}_{-} \leftarrow \min_a \bar{R}^{\text{Tr}}_a(h); \quad R^{\text{Val}}_{-} \leftarrow \min_a \bar{R}^{\text{Val}}_a(h)$
end
Return: $h^*, R^{\text{Val}}_{-}$
```

A.3 Neural Architectures

Table 4 summarizes network architectures and loss functions for all experiments in Section 5. Note that all network heads are EnsembleHeads instead of the standard Dense layer head, these EnsembleHeads were introduced to reduce generalization error, and are described in Section A.3.1.

Dataset	Network Body	Gate	Loss type	Network Head
Synthetic	FullyConnected 64x64	ELU	CategoricalMSE	EnsembleHead 16-head, p=0.5
Adult Dua & Graff (2017a)	FullyConnected 64x64	ELU	CrossEntropy	EnsembleHead 16-head, p=0.5
MIMIC-III Johnson et al. (2016)	FullyConnected 2048x2048	ELU	CrossEntropy	EnsembleHead 16-head, p=0.5
HAM10000 Tschandl et al. (2018)	DenseNet121 Huang et al. (2017)	ReLU	CategoricalMSE	EnsembleHead 16-head, p=0.5

Table 4: Summary of network architectures and losses. All output network heads are 16-head ensemble dense layers to improve generalization errors.

A.3.1 Ensemble Heads

The use of ensemble networks can help reduce the generalization error (Lakshminarayanan et al. (2017)). We found gaps in generalization error to be particularily significant when attempting to bridge risk disparity gaps. To that end, we propose a minimalistic, lightweight implementation of Ensemble heads. The algorithm used to recover the ensemble output at train-time is similar to a structured dropout layer, where several independent dense layers are stochastically averaged to recover the final output, this operation can be efficiently implemented using a single, large dense layer and sampling combination weights as described in Algorithm 3. For test-time evaluation, we instead take the mean across all ensemble heads as the final output.

Algorithm 3: EnsembleHead

```
Given: $x \in \mathbb{R}^{B\times L}$, $n_p, n_o, p, \theta \in \mathbb{R}^{L\times(n_e\times n_0)}$
$x \in \mathbb{R}^{B\times L}$ // Batch of feature vectors
$n_p, n_o$ // Number of ensemble heads and outputs
$\theta \in \mathbb{R}^{L\times(n_e\times n_0)}$ // Parameters of base dense layer
$p$ // Probability of head activity
$y_e = \text{Reshape}(x \times \theta; [B, n_o, n_p])$ // Get full ensemble outputs
$\gamma_e = [\text{Ber}(p)]_{B\times n_p}$ // Get active heads in batch
$y = \frac{y_e \times \gamma_e}{\sum_e \gamma_e}$ // Get mean output of active heads
Return: $y$
```

Ensembles have also been shown to empirically outperform single models, and how they can partially account for risk variance Domingos (2000).

A.4 Analysis of Pareto-optimal classifiers in the infinite data and model capacity regime

In this section we analyze the form of Pareto-optimal solutions to classification (and regression) tasks in the asymptotically ideal case where we have infinite capacity hypothesis classes, that is, or hypothesis class $\mathcal{H}$ contains every function mapping points from the observation space $\mathcal{X}$ to the classification or regression space $\mathbb{R}^{|\mathcal{Y}|}$.

We additionally assume that the joint distributions between target variables Y and observation variables X given every sensitive attribute are known (i.e., $P(Y, X|A)$ is known), and that the loss function $L(h(x), y)$ is convex with respect to $h(x)$.

Lemma 3.2 shows that any joint loss of the form

$$\phi_{P_A}(\{R_a(h)\}) = \mathbb{E}_{X,Y,A}[L(h(X), Y] = \sum_{a\in\mathcal{A}} P_a R_a(h), \tag{5}$$

produces solutions which are already in the Pareto front of $\{R_a(h)\}$ for any distribution of $A \sim P_A$. Since the loss function is convex with respect to $h(X)$ and the hypothesis class is complete, it can also be shown that for any point in the Pareto front, there exists a value of P_A such that that point is reached by minimizing $\phi_{P_A}(\{R_a(h)\})$ Geoffrion (1968); Koski (1985); Miettinen (2012)

We can therefore analyze the Bayes-optimal classifier h_{P_A} which minimizes the Naive risk $\phi_{P_A}(\{R_a(h)\})$, and analytically compute the sub-population risks $R_a(h_{P_A})$ induced. In general, we can write

$$\begin{aligned}
\mathbb{E}[Y|x] \quad &= \sum_{a\in\mathcal{A}} \mathbb{E}[Y|x,a]P(a|x), \\
&= \frac{\sum_{a\in\mathcal{A}} \mathbb{E}[Y|x,a]P(x|a)P_a}{\sum_{a\in\mathcal{A}} P(x|a)P_a}, \\
h_{P_A}(X) \quad &= \arg\min_{\eta} \mathbb{E}[L(\eta, Y)|X], \\
R_a(h_{P_A}) \quad &= \mathbb{E}_{X,Y}[L(h_{P_A}(X), Y)|A = a],
\end{aligned} \tag{6}$$

and in the particular case where target variable Y is categorical, and the classifier loss is an L2 loss against the one-hot encoding of variable Y the equations reduce to

$$\begin{aligned}
\mathbb{E}[Y|x] \quad &= \frac{\sum_{a\in\mathcal{A}} \vec{P}[Y|x,a]P(x|a)P_a}{\sum_{a\in\mathcal{A}} P(x|a)P_a}, \\
h_{P_A}(X) \quad &= \mathbb{E}[Y|x], \\
R_a(h_{P_A}) \quad &= \mathbb{E}_X[\textstyle\sum_y P(y|a,X)\sum_{y'}(h^{y'}_{P_A}(X) - \delta[y-y'])^2].
\end{aligned} \tag{7}$$