Fairness With Minimal Harm:
A Pareto-Optimal Approach For Healthcare

Natalia Martinez[†], Martin Bertran[†], Guillermo Sapiro

Department of Electrical and Computer Engineering
Duke University
{natalia.martinez,martin.bertran,guillermo.sapiro}@duke.edu

Abstract

Common fairness definitions in machine learning focus on balancing notions of disparity and utility. In this work, we study fairness in the context of risk disparity among sub-populations. We are interested in learning models that minimize performance discrepancies across sensitive groups without causing unnecessary harm. This is relevant to high-stakes domains such as healthcare, where non-maleficence is a core principle. We formalize this objective using Pareto frontiers, and provide analysis, based on recent works in fairness, to exemplify scenarios were perfect fairness might not be feasible without doing unnecessary harm. We present a methodology for training neural networks that achieve our goal by dynamically re-balancing subgroups risks. We argue that even in domains where fairness at cost is required, finding a non-unnecessary-harm fairness model is the optimal initial step. We demonstrate this methodology on real case-studies of predicting ICU patient mortality, and classifying skin lesions from dermatoscopic images.

1   Introduction

Machine learning algorithms play an important role in decision making in society. When these algorithms are used to make high-impact healthcare decisions such as deciding whether or not to classify a skin lesion as benign or not, or predicting mortality for intensive care unit patients, it is paramount to guarantee that these decisions are 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 (Barocas and Selbst [2016]). It is desirable in these critical applications to impose some fairness criteria. There are several lines of work in fairness in machine learning, such as Demographic Parity (Louizos et al. [2015], Zemel et al. [2013], Feldman et al. [2015]), Equality of Odds, Equality of Opportunity (Hardt et al. [2016], Woodworth et al. [2017]), or Calibration (Pleiss et al. [2017]). These notions of fairness are 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 to any subgroup (Ustun et al. [2019]).

In this work, we measure discrimination (unfairness) in terms of difference in predictive risks across sub-populations defined by our sensitive attributes. This concept has been explored in other works such as Calders and 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 contrast to common post-hoc correction methods like the ones proposed in Hardt et al. [2016], Woodworth et al. [2017] where randomness is potentially added to the decisions of all sub-populations. While this type of approach diminishes the accuracy-disparity gap, it does so by potentially introducing randomness into the final decision (i.e., with some probability, disregard classifier output and produce arbitrary decision based solely on sensitive label), which leads to performance degradation in most common risk metrics. Since our proposed methodology does not

---

† Equal contributions

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-unnecessary-harm fairness using Pareto optimality, 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 is both Pareto optimal with respect to the sub-population risks (no unnecessary harm) and minimizes risk disparity. This notion is already amenable to non-binary sensitive attributes. We analyze Pareto fairness on an illustrative example, and compare it to alternative approaches. We provide an algorithm that promotes fair solutions belonging to the Pareto front; this algorithm can be applied to any standard classifier or regression task. Finally, we show how our methodology performs on real tasks such as predicting ICU mortality rates in the MIMIC-III dataset from hospital notes Johnson et al. [2016], and classifying skin lesions in the HAM10000 dataset Tschandl et al. [2018].

2   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 our 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. To formalize this notion, we define:

Definition 2.1. Pareto front: The set of Pareto front classifiers is defined as $\mathcal{P}(\mathcal{H}, \mathcal{A}) = \{h \in \mathcal{H} :/ \exists h' \in \mathcal{H}, \nexists a' \in \mathcal{A} : [R_{a'}(h') < R_{a'}(h)] \wedge [R_a(h') \leq R_a(h)], \forall a \in \mathcal{A}\}$

Definition 2.2. Pareto-fair classifier and Pareto-fair vector: A classifier h^* is an optimal noharm 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 vector is defined as $\boldsymbol{r}^* \in \mathbb{R}^{|\mathcal{A}|} : \boldsymbol{r}^* = \{R_a(h^*)\}_{a \in \mathcal{A}}$.

The Pareto front defines the best achievable trade-offs between population risks $R_a(h)$, while the Pareto-fair classifier gives the trade-off with least disparity. Building on analysis done by Chen et al. [2018], Domingos [2000], it is possible to decompose the risk in bias, variance and noise for some given loss functions. The noise represents the smallest achievable risk for infinitely large datasets. If it differs between sensitive groups, zero-discrimination (perfect fairness) can be only achieved by introducing bias or variance, hence doing harm. Figure 1 shows a scenario where the Pareto-front does not intersect the equality of risk line for the case of two sensitive groups $a \in \{0, 1\}$ and a binary output variable $y \in \{0, 1\}$. Here the level of noise between subgroups differs, and the Pareto-fair vector $\boldsymbol{r}^*$ would not be achieved by either a naive classifier (minimizes expected global risk), or a classifier where low-occurrence subgroups are over-sampled (re-balanced naive classifier).

3   Methods

Any loss function that is monotonically decreasing with $R_a(h), \forall a$ is minimized by a classifier in the Pareto front. Since we want to minimize the discrimination gap, we will build an adaptive loss function that is shares this property with the following form:

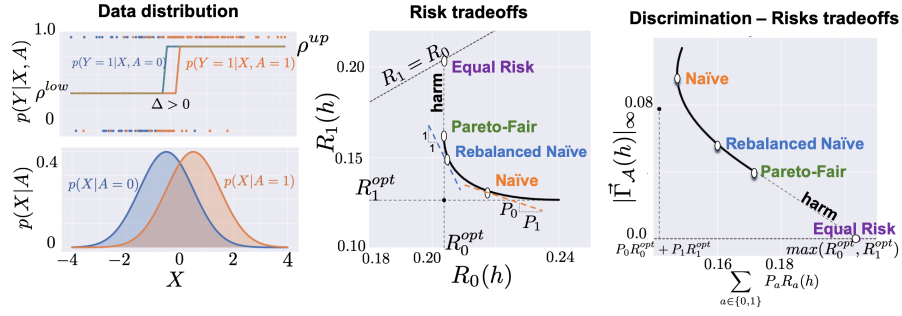$$\phi(h; \boldsymbol{\mu}, c) \quad = \sum_{a \in \mathcal{A}} R_a(h) + \mu_a(R_a(h) - c)^{+2}, \tag{1}$$

Figure 1: Example of binary target (y) and sensitive (a) variable. Bottom left figure shows conditional distributions of observation variable (x) $p(x|a)$, while top left shows distribution of y as a function of $x, a, p(y|x, a)$, these distributions are simple piece-wise constants with two levels ρ^{low} and ρ^{high}, Δ represents the gap between level transitions for each group. Middle image shows the Pareto front between group risks, 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 Re-balanced Naive classifiers do not attain the best possible no-harm classifier. Pareto-fair point is shown in green. Rightmost figure shows the trade-offs attainable between discrimination and mean risks, an equivalent problem to the risk trade-off figure.

with $c < \min_a R_a(h)$ and $\boldsymbol{\mu} = \{\mu_a\}_{a \in \mathcal{A}}$. It can be shown that for convex Pareto sets (with respect to risk vectors $\boldsymbol{r} = \{R_a(h)\}_{a \in \mathcal{A}}$), there exist a set of $\boldsymbol{\mu}^*$ such that the Pareto-fair vector $\boldsymbol{r}^*$ is a unique solution. In Algorithm 1 we jointly search for $\boldsymbol{\mu}^*$, $\boldsymbol{r}^*$, and h^*.

Algorithm 1: ParetoFairOptimization

Given: $h_\theta, \mathcal{D}, \gamma > 0, \xi \in (0, 1), \zeta \in (0, 1), \text{lr},$
$\boldsymbol{\mu} \leftarrow \mathbf{0}, \ \boldsymbol{\mu}^* \leftarrow \mathbf{0}, \ , c \leftarrow 0, \ h^* \leftarrow h_\theta$
while *Improving* **do**

    $h_\theta \leftarrow \text{SGDwithEarlyStopping}(h_\theta, \phi, \boldsymbol{\mu}, c, \mathcal{D})$ // Optimize adaptive loss with fixed $\boldsymbol{\mu}$ on dataset $\mathcal{D}$

    $r^{\text{val}} \leftarrow \text{EvaluateRisk}(h_\theta, \mathcal{D}^{\text{val}})$

    if $||\vec{\Gamma}_{\mathcal{A}}(h)||_\infty < \Gamma^*$ *and* $\boldsymbol{r}^{Val}$ *is not dominated by previous validation risks* **then**

        $h^* \leftarrow h_\theta, \ \Gamma^* \leftarrow ||\vec{\Gamma}_{\mathcal{A}}(h)||_\infty, \ c_{\text{old}} \leftarrow c, \ c \leftarrow \min_a \frac{r_a^{\text{Val}}}{k}$

        $\boldsymbol{\mu}^* \leftarrow \boldsymbol{\mu} \cdot \frac{(\boldsymbol{r}^{\text{Val}} - c_{\text{old}})^+}{(\boldsymbol{r}^{\text{Val}} - c)^+}, \ a' \leftarrow \arg\max_a r_a^{\text{Val}}$

    else

        $\text{lr} \leftarrow \zeta \, \text{lr}, \boldsymbol{\mu} \leftarrow \boldsymbol{\mu}^*, \ \gamma \leftarrow \gamma\xi, \ h_\theta \leftarrow h^*$

    end

    $\mu_{a'} \leftarrow (1 + \gamma)\mu_{a'}$

end

Return: h^* // Exit loop due to excessive iterations or no improvement in fairness

---

4 Experiments and Results

We evaluate our method on mortality prediction and skin lesion classification and show empirically how we reduce accuracy and risk disparity without unnecessary harm. We compare it against a naive classifier, class-rebalancing (equally sampled sub-groups), the post-processing framework presented in Hardt et al. [2016], and the Disparate Mistreatment framework of Zafar et al. [2017].

4.1 Predicting Mortality in Intensive Care Patients

We analyze clinical notes 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] and use tf-idf statistics on the $10,000$ most frequent words in clinical notes as input features. Fairness is measured with respect to age (under/over 55 years old), ethnicity, and outcome. We used a fully connected neural network with two 2048-unit hidden

layers trained with Brier score (BS) loss. Table 1 shows accuracy and BS of all tested methodologies. We observe from columns (PF BS) and (PF Acc) that our model has the best accuracy and BS discrepancies. We can reduce accuracy disparities further by applying Hardt post processing (HPF Acc), at the cost of reducing overall performance. Note that, as expected, it is better to apply this post processing on our method than on the rebalanced naive classifier (HPF Acc vs HRen Acc). This illustrates the goal of the proposed Pareto-fair paradigm: develop the fairest algorithm with no unnecessary harm, and if (e.g., due to policy) the resulting fairness level needs to be improved, use other post-processing techniques on the Pareto-fair classifier such as Hardt et al. [2016].

Out/Age/Race	Ratio	Na Acc	ReN Acc	Zafar Acc	PF Acc	HReN Acc	HPF Acc	ReN BS	PF BS
A/A/NW	5.7%	99.1±0.4%	86.3±1.5%	93.0±1.4%	83.4±2.6%	76.3±1.9%	71.6±3.2%	0.2 ± 0.02	0.25 ± 0.03
A/A/W	13.3%	98.8±0.5%	86.3±1.1%	90.0±1.3%	83.2±1.5%	76.7±1.6%	71.8±1.9%	0.2 ± 0.01	0.25 ± 0.02
A/S/NW	12.9%	97.5±0.6%	76.5±1.7%	81.8±1.7%	71.4±3.0%	76.4±2.2%	71.3±3.1%	0.31 ± 0.02	0.36 ± 0.03
A/S/W	56.7%	97.9±0.3%	79.0±0.6%	77.4±0.7%	74.6±1.6%	76.2±2.2%	72.1±1.2%	0.28 ± 0.01	0.34 ± 0.02
D/A/NW	0.4%	23.4±9.4%	76.1±8.5%	47.7±9.6%	78.6±6.1%	66.6±9.9%	74.1±9.2%	0.36 ± 0.06	0.34 ± 0.04
D/A/W	0.9%	32.6±3.7%	80.1±3.3%	60.5±6.3%	83.3±4.2%	66.4±2.4%	73.6±4.3%	0.29 ± 0.04	0.28 ± 0.03
D/S/NW	1.8%	21.4±2.2%	66.9±2.4%	48.2±2.0%	73.3±2.9%	64.8±2.1%	71.2±2.9%	0.42 ± 0.02	0.37 ± 0.03
D/S/W	8.3%	23.4±2.2%	67.4±1.9%	57.1±2.2%	72.5±3.6%	66.2±2.9%	72.4±3.5%	0.42 ± 0.02	0.37 ± 0.04
Sample Mean	-	**89.5±0.2%**	78.9±0.7%	78.0±0.7%	75.7±1.1%	**75.1±1.8%**	71.9±1.2%	**0.29 ± 0.01**	0.33 ± 0.01
Group Mean	12.5%	61.8±1.5%	77.3±1.3%	69.4±1.3%	**77.5±0.7%**	71.2±1.1%	**72.3±1.0%**	**0.31 ± 0.01**	0.32 ± 0.01
Discrepancy	56.3%	81.1±2.5%	22.5±2.5%	49.3±4.4%	**17.1±2.6%**	18.6±3.2%	**13.1±5.6%**	0.24 ± 0.03	**0.16 ± 0.03**

Table 1: Result summary table on MIMIC dataset. Sensitive attributes are the combination of Outcome (Alive or Deceased), Age (Adult or Senior) and Race (White or Non-White). Target attribute is Outcome. We report Brier scores (BS) and Accuracies (Acc) for Naive (Na), Rebalanced Naive (ReN), Zafar (Zafar), and Pareto-Fair (PF) classifiers; the prefix (H) indicates that Hardt post-processing was applied to control for Age and Race. Zafar was applied on top of the embeddings of the Naive classifier, as that produced the best results.

4.1.1 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 in 7 categories. We used a pretrained DenseNet121 Huang et al. [2017] as our base classifier, and measured fairness with respect to diagnosis class, casting balanced risk minimization as a particular use-case for Pareto fairness. Table 2 shows our empirical results.

Groups	Ratio	PF Acc	ReN Acc	Na Acc	PF BS	ReN BS	Na BS
akiec	2.5%	51.9%	55.6%	3.7%	0.741	0.671	1.289
bcc	2.7%	56.7%	76.7%	56.7%	0.549	0.341	0.613
bkl	7.2%	59.5%	58.2%	36.7%	0.62	0.606	0.931
df	0.5%	66.7%	33.3%	0.0%	0.536	0.898	1.721
nv	81.0%	83.5%	90.9%	96.0%	0.241	0.128	0.054
vasc	1.3%	71.4%	85.7%	0.0%	0.36	0.246	1.604
mel	4.8%	53.8%	48.1%	32.7%	0.586	0.657	0.941
Sample Mean	-	78.6%	**84.8%**	83.6%	0.308	**0.213**	0.234
Group Mean	14.3%	63.4%	**64.1%**	32.3%	0.519	**0.507**	1.022
Discrepancy	80.4%	**31.7%**	57.5%	96.0%	**0.501**	0.769	1.667

Table 2: Group-specific risks and accuracies for Naive (Na), Rebalanced Naive (ReN), and Pareto-Fair (PF) classifiers. Lesion Type was chosen as both target and sensitive variable, they 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 exhibits the lowest discrepancy on Acc and BS while still being Pareto-optimal.

5 Discussion

Here we explore the problem of reducing the risk disparity gaps in the most ethical way possible (i.e., minimizing unnecessary harm). We provide an algorithm that can be used with any standard classification or regression loss to bridge risk disparity gaps without introducing unnecessary harm. We show its performance on two real-world case studies; we take advantage of the fact that our method does not require test-time access to sensitive attributes to frame balanced classification as a fairness problem. In future work, we wish 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 believe that no-unnecessary-harm notions of fairness are of great interest for several applications, especially so on domains such as healthcare.

References

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*, pages 3539–3550, 2018.

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

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*, pages 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*, pages 259–268. ACM, 2015.

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

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*, pages 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.

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

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.

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*, pages 6373–6382, 2019.

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 Rodriguez, Krishna Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pages 229–239, 2017.

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