Pareto-Efficient Fairness for Skewed Subgroup Data

Ananth Balashankar [1 2 *] **Alyssa Lees** [2] **Chris Welty** [2] **Lakshminarayanan Subramanian** [1]

Abstract

As awareness of the potential for learned models to amplify existing societal biases increases, the field of ML fairness has developed mitigation techniques. A prevalent method applies constraints, including equality of performance, with respect to subgroups defined over the intersection of sensitive attributes such as race and gender. Enforcing such constraints when the subgroup populations are considerably skewed with respect to a target can lead to unintentional degradation in performance, without benefiting any individual subgroup, counter to the United Nations Sustainable Development goals of reducing inequalities and promoting growth. In order to avoid such performance degradation while ensuring equitable treatment to all groups, we propose Pareto-Efficient Fairness (PEF), which identifies the operating point on the Pareto curve of subgroup performances closest to the fairness hyperplane. Specifically, PEF finds a Pareto Optimal point which maximizes multiple subgroup accuracy measures. The algorithm *scalarizes* using the adaptive weighted metric norm by iteratively searching the Pareto region of all models enforcing the fairness constraint. PEF is backed by strong theoretical results on discoverability and provides domain practitioners finer control in navigating both convex and non-convex accuracy-fairness trade-offs. Empirically, we show that PEF increases performance of all subgroups in skewed synthetic data and UCI datasets.

1. Introduction

Increased awareness of the potential for machine learning models to amplify existing societal biases (Bolukbasi et al., 2016) has informed work in AI systems. Several approaches to mitigate bias rely on defining a set of sensitive variables whose combinations of values create subgroups of data, often with different real-world distribution characteristics, such as skew. However a known trade-off between fairness constraints and classifier accuracy exists (Menon & Williamson, 2018). We propose a fairness constraint based on Pareto-Efficiency (Godfrey et al., 2007) to avoid unintentional degradation in subgroup performance, while striving for increased accuracy. The Pareto-Efficient Fairness (PEF) constraint restricts the choice of ML models to the Pareto frontier to ensure higher accuracy across *fair* model options.

Existing fairness mitigation algorithms often explicitly define constraints on model subgroup performance (e.g Equality of Odds (Hardt et al., 2016)), enforced using Lagrangian relaxation (Menon & Williamson, 2018; Liu et al., 2018; Burke, 2017). However, (Menon & Williamson, 2018) has established that such approximate group fairness constraints are not *perfectly* possible unless the underlying subpopulations demonstrate perfect accuracy with respect to the target. This inherent tradeoff between accuracy and fairness is often amplified due to sampling biases in the underlying sub-population distributions. The skew in sub-population data distributions, due to underlying population traits, varied subgroup population prevalence, sampling bias, etc, will be propagated in a model if not addressed (Zhao et al., 2017).

Similarly, as the number of sensitive variables increase, real-world sub-populations at the intersection of variables may be extremely limited to sample from (Kearns et al., 2017). Forcing subgroup metric constraints in these use-cases may result in trivial accuracy (equivalent to a random coin toss), since equivalency metrics are constrained by the model's worst performing subgroup (Menon & Williamson, 2018). The increase in sensitive variables remains a critical problem in real case studies, where existence of highly correlated proxy variables expand the set of variables to be considered (Bickel et al., 1975; Chiappa & P. S. Gillam, 2018; Dheeru & Karra Taniskidou, 2017).

Our contribution of the Pareto-Efficiency Fairness (PEF) criterion selects a model whose individual subgroup performance exceeds that of all other models. In some cases, a Pareto-Efficient definition may be at odds with a strict equality fairness criterion. Figure 1 illustrates cases where

[*]Work done during an internship at Google AI [1]New York University, New York, USA [2]Google AI, New York, USA. Correspondence to: Ananth Balashankar <ananth@nyu.edu>.

Appearing at the International Conference on Machine Learning AI for Social Good Workshop, Long Beach, United States, 2019.

extremely unfair models might be Pareto optimal and vice versa. However, our algorithm avoids this pitfall by limiting the search space on the Pareto frontier within given fairness bounds.

Our proposed bias loss function achieves Pareto-Efficient performance, which by definition is superior in performance to solutions based on equalizing subgroup performance. Based on theory of multiple objective optimization for continuous Pareto fronts, we show that our approach achieves an operating point which is better both in terms of global accuracy and individual subgroups accuracy than methods which approximate hard constraints of equality (Zhao et al., 2017) and adversarial multi-task learning (Beutel et al., 2017) on both synthetic data and UCI datasets.

2. Related Work

Much previous work has explored methodologies for achieving subgroup fairness in ML classification. (Zhao et al., 2017) aims to achieve corpus level parity with Lagrangian relaxation. Updates after each batch of training are approximations on samples with the goal of achieving corpus level parity across sensitive variables. The work has not been extended to a large number of subgroups.

(Menon & Williamson, 2018) show that a disparate impact constraint is equivalent to a cost sensitive constraint. The work formulates a fairness frontier: for a given lower bound on fairness, they calculate the best excess risk over the solution without a fairness constraint, along with a data dependent theoretical limitation between fairness and accuracy. We extend on these results and show that if the fairness frontier is steep, then PEF achieves better efficiency than existing constrained optimization methods.

(Beutel et al., 2017) models the problem of debiasing as a multi-task learning problem where the model is penalized if the shared hidden layers of a neural network can be used to predict the sensitive variable accurately. One potential issue of the model is that it could result in propagating bias in the reverse direction.

There are many fairness definitions proposed in literature, but we address one such definition in this work, Equality of Odds. (Hardt et al., 2016) We say that a predictor $\hat{Y}$ satisfies equalized odds with respect to protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y.

$$P(\hat{Y} = \hat{y}|Y = y, A = m) = P(\hat{Y} = \hat{y}|Y = y, A = n)$$
$$\forall Y, \forall m, n \in A$$

(Pleiss et al., 2017) and (Raghavan et al., 2018) prove that equality of odds can't be achieved by two models on separate groups which are calibrated , unless both the models achieve perfect accuracy. The main intuition behind this paper is the hypothesis that similar impossibility regimes exist in real life scenarios, especially when multiple subgroups exist. We explore how to avoid unintentional performance degradation in such cases, which only achieve fairness by trivially reducing accuracy to random performance for all subgroups.

Pareto Efficiency is a state where resources are allocated in which it is impossible to redistribute resources to make any one criterion or party better off without making another criterion worse off.

We seek to mitigate some of the weaknesses described above by employing Pareto-Efficiency or Pareto Optimality. With respect to fairness, (Agarwal et al., 2018) explores the Pareto optimality between overall accuracy and violation of fairness constraints. Although such a comparison is important and in many cases necessary by a domain expert, it is a measure of two separate metrics and needs to be carefully evaluated. However in our work, we focus on the trade-offs between the performance of various comparable subgroups that can be easily compared simultaneously on the Pareto-optimal curve. To the best of our knowledge, this is the first work which extends strong theoretical results of Pareto-Efficiency to achieve better subgroup performance in data distributions with high disalignment between fairness and accuracy.

3. Methodology

3.1. Overview

In order to illustrate the benefits of Pareto-Efficient Fairness, consider a binary classification task $X \rightarrow \{0, 1\}$, where X is a continuous scalar feature. We define a set of two groups $G = \{A, B\}$ with membership over a sensitive variable set S. Figure 1 shows the scatter plot of achievable accuracy metrics over groups A and B for classifiers h ∈ H of the form

$$h(X) = \begin{cases} 0, & \text{if } X < t \\ 1, & \text{otherwise} \end{cases} \quad (1)$$

by varying values of t. The expected accuracy for a uniformly random population (shown as the red point in Figure 1) over a class-balanced set of both groups A and B is .5, if the target label is predicted by flipping a fair coin. We denote the best *known* accuracies over all operating points for A and B, given the set of classifiers H, as opt_a and opt_b. Equality of odds would require that the classifier operate on the $x = y$ line.

However, if the objective is to improve the performance of protected groups to the meet the levels of the highest performing groups as motivated by policies around affir-

mative action (Foster & Vohra, 1992) and recent works in fairness literature (Buolamwini & Gebru, 2018), then choosing points on the line $x = y$ might not be desirable. Instead, choosing one of the Pareto-Efficient points (blue) may be a more desirable solution in that it will increase the accuracy of **both** groups A and B compared to a solution using equalized odds.

In our employment of Pareto-Efficient subgroup fairness, we define the Pareto-Efficient points as *the set of points for which there does not exist another point, which is better performing across all the sensitive groups.*

One possible pitfall of such a definition is that points opt_a or opt_b could be selected as a Pareto-Efficient solution. Both are trivially Pareto-Efficient points, since there are no other points which performs better across all groups, but clearly are unequal across subgroups. In order to avoid disparity in the Pareto-Efficient Fairness penalty across subgroups, we minimize the variance of penalty across subgroups while choosing among the Pareto-Efficient points. Formally, for a set of Pareto-Efficient thresholds: t_{PE}, performance metric for g: $F[g]$ and optimum performance metric across all operating points for group g: $F_{opt}[g]$, we intend to find,

$$t_{fair} = \min_{t_{PE}} \sigma_g^2(\epsilon_g) \tag{2}$$

$$\text{where } \epsilon_g = 1 - \frac{F[g]}{F_{opt}[g]}, \sigma_g \text{ is the stddev across groups} \tag{3}$$

We would choose a Pareto-Efficient threshold empirically using the minimization criteria,

$$t_{PE} = \min_{t} \|\epsilon_g\|_1 \tag{4}$$

These two minimization criterion might not concur, hence we need a Lagrangian approximation which combines the two as follows:

$$t_{PE-fair} = \min_{t} \alpha\|\epsilon_g\|_1 + (1-\alpha)\sigma_g^2(\epsilon_g) \tag{5}$$

The choice of α is domain dependent with the nuanced point that α is defined relative to the respective heuristic pseudo-optimum performance of groups, which is central to the argument of Pareto-Efficient Fairness. This is demonstrated when $\alpha = 0$ and the variance of the pareto-errors ϵ_g is minimized, which is not the same as equality of odds. Similarly, when $\alpha = 1$, we minimize the sum of absolute pareto-errors, which is a different formulation than the unconstrained optimization in (Zhao et al., 2017)

These criterion can be used as a regularizer in a Lagrangian dual formulation similar to (Eban et al., 2016) with the appropriate loss weight (λ), along with a cross entropy classification loss: $L_{ce}(o, \hat{o})$ to yield

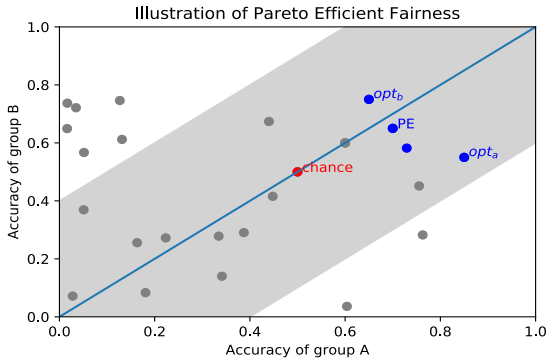

Figure 1. Illustration of Pareto Efficiency Fairness within the fairness relaxation bounds (in gray) in a dataset with 2 groups. A scatter plot of subgroup level accuracies that are achievable by models are plotted here. The line denotes permissible accuracies when strict equality constraints are enforced. The most accurate fair model has subgroup accuracies of (0.6, 0.6). If accuracy for each of the individual groups is separately maximized, we would choose points $opt_a = (0.83, 0.55)$, and $opt_b = (0.63, 0.77)$ respectively. However, PEF will choose among the points on the Pareto-front (in blue), the point $PE = (0.71, 0.63)$, which minimizes variance of the Pareto loss across subgroups.

$$L_p(o, \hat{o}) = L_{ce}(o, \hat{o})) + \sum_{g \in G} (\lambda(\alpha\|\epsilon_g\|_1 + (1-\alpha)\sigma_g^2(\epsilon_g))$$

3.2. Pareto-Efficient Algorithm

In order to obtain a heuristic for the pseudo-optimal subgroup performance, we obtain a baseline metric based on the performance of the same classifier set trained purely on individual subgroup data. Although this may not be the best estimate of a subgroup population's true optimal, we avoid transfer learning during initialization as majority subgroups benefit from a large dataset and minority subgroups can be penalized (Papernot et al., 2016). Instead, we propose an iterative approach where the Pareto-Loss heuristic is updated in each iteration if the optimal subgroup performance was bettered by a jointly trained model. In the evaluation, we present results from the heuristic initialized with the separately trained subgroups. Once we obtain the heuristic pseudo-optimal performances, we input them into the jointly trained Pareto bias mitigation algorithm (Algorithm 1), which minimizes the loss L_p mentioned above for every batch. We further ensure that the batch used at every loss minimization is representative of the subgroup distribution present in the original dataset.

Our work introduces the notion of a "potentially optimal" performance for each of the sub-population subgroups. We

Algorithm 1 Iterative Pareto-Efficient Bias Mitigation

G = $\mathcal{P}(S)$ where S is the membership set of sensitive variables
$F_{opt}(g) = eval(M_g, D_g)\ \forall g \in G$, M_g is the model trained and evaluated on only data with subgroup g: D_g
$F = \{\}$
while F is $\{\}$ or $\exists g \in G, F[g] > F_{opt}[g]$ **do**
  Train M to minimize L_p for the updated F_{opt}
  $F[g] = eval(M, D_g), \forall g \in G$
  Update $F_{opt}[g] = F[g]$ if $F[g] > F_{opt}[g]$
end while

Table 1. UCI Adult dataset with bias mitigation algorithms

Model	Accuracy	FPR	FNR	Discrepancy	Pareto Loss
Baseline (no bias loss)	0.630	0.253	0.747	0.199	0.016
Minimize Discrepancy	0.619	0.283	**0.712**	**0.167**	0.133
Adversarial Loss	0.648	0.224	0.769	0.226	0.077
Pareto-Efficient Loss	**0.678**	**0.165**	0.830	0.250	**0.000**

Table 2. Subgroup performance on UCI Adult dataset

Model	Subgroup 1	2	3	4	Pareto Loss
Baseline (no bias loss)	0.890	0.883	0.818	0.784	0.016
Minimize Discrepancy	0.853	0.856	0.806	0.778	0.133
Adversarial Loss	0.882	0.872	0.824	0.780	0.077
Pareto-Efficient Loss	**0.935**	**0.915**	**0.844**	**0.797**	**0.000**
Subgroup Pareto Frontier	0.934	0.894	0.815	0.783	N/A

argue that at a minimum a classifier aiming towards *fairness* should be able to reflect the underlying sampled subpopulations distributions as accurately as possible. For example, in a synthetic data distribution, this *pseudo-optimal* performance can be calibrated differently for each subgroup as seen in Table 2. In this scenario, there are subgroups that do not perform better, independent of the threshold chosen, whereas there are two other subgroups that perform better for higher values of the threshold. Hence, we achieve equal performance when all subgroups perform equal to random chance. But, choosing to minimize Pareto loss is better as those Pareto-optimal points *dominate* the result which minimizes the parity error across all subgroups.

This approach is similar to current avenues of research which highlight the benefits of fairness through awareness. In this interpretation of fairness, it is acceptable to understand the differences between various subgroups' performance in the dataset and operate in a way to correct as opposed to fairness through blind application.

4. Evaluation

We compare our approach with the scaled versions of group fairness (Zhao et al., 2017) and (Beutel et al., 2017) for subgroups. In (Zhao et al., 2017), the authors optimize for overall accuracy while constraining for equality across false positive rates, but the method is also applicable to other measures of performance. Our implementation employs Lagrangian relaxation to add a penalty for each subgroup that deviates from the overall accuracy. In (Beutel et al., 2017), Beutel et al, the authors implement bias mitigation as a way of erasing the sensitive group membership by back-propagating negative gradients in a multi-headed feed-forward neural network. We scale the same for subgroups defined over multiple sensitive variables, where the auxiliary head aims to predict the subgroup class (multi-class classification instead of binary). We evaluate a comparison of the techniques for both synthetic toy data and the UCI Census Adult dataset. The UCI Census Adult dataset predicts income category based on demographic information, where the sensitive variables selected for experiments are set as gender and race.

4.1. UCI Census Data

Table 1 shows the Pareto-loss, i.e how much each subgroup deviates from the pseudo-optimal of the respective subgroup for the UCI Census Adult dataset. We see that our approach achieves zero Pareto-Loss, while (Zhao et al., 2017) and (Beutel et al., 2017) have non-zero Pareto losses. (Zhao et al., 2017) performs well in terms of lowering the sum of absolute discrepancy of all subgroups' accuracy from the overall accuracy. This is expected as (Zhao et al., 2017) chooses an operating point closest to equal accuracy, when exact equality isn't possible. (Beutel et al., 2017) arrives at an operating point which suffers from non-zero discrepancy and Pareto-loss. Table 2 clarifies why our approach arrives at a better operating point. We can see that each of the subgroups have better individual accuracy than all the other approaches, some even better than the baseline. This confirms empirically that our objective function matches (and sometimes exceeds due to transfer learning) the heuristic pseudo-optimal performance for each subgroup (quoted in the last row of Table 2).

5. Conclusion

In this paper, we establish Pareto-Efficient Fairness over subgroups improves overall accuracy as well as subgroup performance metrics on synthetic and UCI Adult datasets. Based on theoretical results, we show that PEF is more efficient in datasets with high skew and converges to a Pareto Optimal point which dominates all existing methods which enforce a strict constraint.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. M. A reductions approach to fair classification. *CoRR*, abs/1803.02453, 2018. URL

http://arxiv.org/abs/1803.02453.

Beutel, A., Chen, J., Zhao, Z., and hsin Chi, E. H. Data decisions and theoretical implications when adversarially learning fair representations. *CoRR*, abs/1707.00075, 2017.

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. Sex bias in graduate admissions: Data from berkeley. *Science*, 187(4175):398–404, 1975. ISSN 0036-8075. doi: 10.1126/science.187.4175.398. URL http://science.sciencemag.org/content/187/4175/398.

Bolukbasi, T., Chang, K., Zou, J. Y., Saligrama, V., and Kalai, A. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520, 2016. URL http://arxiv.org/abs/1607.06520.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A. and Wilson, C. (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR. URL http://proceedings.mlr.press/v81/buolamwini18a.html.

Burke, R. D. Multisided fairness for recommendation. *CoRR*, abs/1707.00093, 2017.

Chiappa, S. and P. S. Gillam, T. Path-specific counterfactual fairness. 02 2018.

Dheeru, D. and Karra Taniskidou, E. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Eban, E. E., Schain, M., Mackey, A., Gordon, A., Saurous, R. A., and Elidan, G. Scalable Learning of Non-Decomposable Objectives. *ArXiv e-prints*, August 2016.

Foster, D. and Vohra, R. An economic argument for affirmative action. *Rationality and Society*, 4(2):176–188, 1992. doi: 10.1177/1043463192004002004. URL https://doi.org/10.1177/1043463192004002004.

Godfrey, P., Shipley, R., and Gryz, J. Algorithms and analyses for maximal vector computation. *The VLDB Journal*, 16(1):5–28, January 2007. ISSN 1066-8888. doi: 10.1007/s00778-006-0029-7. URL http://dx.doi.org/10.1007/s00778-006-0029-7.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016. URL http://arxiv.org/abs/1610.02413.

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

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. Delayed impact of fair machine learning. *CoRR*, abs/1803.04383, 2018. URL http://arxiv.org/abs/1803.04383.

Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 107–118, New York, NY, USA, 23–24 Feb 2018. PMLR.

Papernot, N., Abadi, M., Erlingsson, Ú., Goodfellow, I. J., and Talwar, K. Semi-supervised knowledge transfer for deep learning from private training data. *CoRR*, abs/1610.05755, 2016. URL http://arxiv.org/abs/1610.05755.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. M., and Weinberger, K. Q. On fairness and calibration. *CoRR*, abs/1709.02012, 2017. URL http://arxiv.org/abs/1709.02012.

Raghavan, M., Slivkins, A., Vaughan, J. W., and Wu, Z. S. The externalities of exploration and how data diversity helps exploitation. *CoRR*, abs/1806.00543, 2018.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017.

Supplementary Material: Pareto-Efficient Fairness for Skewed Subgroup Data

Ananth Balashankar [1 *] **Alyssa Lees** [2] **Chris Welty** [2] **Lakshminarayanan Subramanian** [1]

1. Properties of Pareto-Efficient Fairness

As noted in Section 3, let the PEF loss over M subgroups be defined as $L_p(o, \hat{o}) = L_{ce}(o, \hat{o}) + \lambda \sum_{i=0}^{M} (\alpha \|L(i)\|_1 + (1-\alpha)\|L(i) - \epsilon\|_2)$, where o, $\hat{o}$ denotes the true and predicted target variable, L_{ce} denote the cross entropy loss, $L(i)$ denote the pareto loss measured for group i, λ and α are hyperparameters which control the tradeoff.

Let $H(\tau)$ denote the disalignment between the target and the subgroup for the fairness constraint, it is given that $H(\tau) = E_X[(c - \eta(X)).(f^*(X) - [[\eta(X) > c)]])]$ where c is the threshold for classification, $\eta(X)$ is the class probability function for the target and f^* denote the Bayes-optimal classifier which minimizes the fairness constrained loss function.

If $H(\tau) = 1 - \delta$, for $\delta \to 0$, then the PEF loss L_p will converge to a Pareto-optimal point of accuracies $\hat{F} = [\hat{f}_1, \hat{f}_2, ..\hat{f}_M]$ for M subgroups, such that for all accuracies obtained by strictly enforcing the fairness constraint, $F = [f_1, f_2, ..f_M]$, we have $\hat{f}_i \geq f_i$.

In the remainder of the section, we outline key results about the PEF algorithm's convergence properties, its capacity to discover Pareto curves of subgroup performances and pareto efficiency that are part of the theorem.

1.1. Convergence

We first present the lemma derived for sparse lasso regularizers (Vincent & Hansen, 2014), that show the convexity of block regularized minimizers, which also minimize the loss function under model parameter (β) constraints. If f is a convex, twice-differentiable loss function, then the sparse lasso minimizer, min(f + $\lambda\phi$), such that $\phi(\beta) = (1-\alpha)\|\beta\|_2 + \alpha|\beta|$, is also convex.

Moreover, it has be shown that the convexity argument holds even when $\beta \in R^n$, as long as the block separability of β holds, i.e = $\phi(\beta) = \sum_{i=0}^{M} \phi^{(i)}(\beta^{(i)})$, where i^{th} component of β denotes the i^{th} subgroup's performance's deviation from their group optimal performance. Hence, adopting a block level gradient descent, where the gradients are back-propagated only after each batch's block performances are computed has been shown to converge in (Tseng & Yun, 2009).

1.2. Discoverability

The above section shows that the Pareto-Efficient Fairness loss indeed converges to a minima with the use of a convex loss function. However, it remains to be seen that the minima obtained is a pareto-optimal operating point. For this, we now provide insights behind the choice of the regularizers in Pareto-Efficient Fairness, based on the theory of multiple objective optimization (Giagkiozis & Fleming, 2012). Specifically, we use the theory of decomposition based methods which employ a scalarization technique to convert M multiple objectives - $f_1.f_2..f_M$ into a single objective using a Weighted Metric method. Here, the distance of each objective from a Utopian reference point is measured and a corresponding l_p norm is minimized, i.e. $min(\left(\sum_{m=1}^{M} w_m |f_m(x) - z_m^*|^p\right)^{\frac{1}{p}})$
The knowledge of a Utopian reference point z^* is usually based on prior domain knowledge. In our adaptation for Pareto-Efficient Fairness, we have initialized z^* to a vector of subgroup performances when trained exclusively on the subgroup's data alone.

Using the above Weighted Metric method provides the ability to discover both convex and non-convex pareto curves as shown in (Giagkiozis & Fleming, 2012). Similarly, the l_∞ norm has the ability to discover all points on the Pareto front for some weight vector as stated in the lemma below. (Miettinen, 1999)

Let x be a Pareto-optimal solution, then there exists a positive weighting w vector such that x is a solution of the weighted Tchebycheff problem
$min(\left(max_{m=1}^{M} w_m |f_m(x) - z_m^*|\right))$, where the reference point z^* is the utopian objective vector.

However, it can be seen that as $p \to \infty$, the objective function becomes non-differentiable and hence it is of interest to us that we choose the minimum possible p for which the above statement still holds. While a universal result for all

*Work done during an internship at Google AI [1]New York University, New York, USA [2]Google AI, New York, USA. Correspondence to: Ananth Balashankar <ananth@nyu.edu>.

Appearing at the International Conference on Machine Learning AI for Social Good Workshop, Long Beach, United States, 2019.

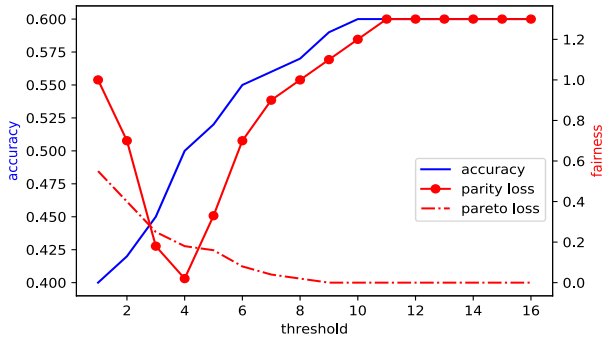

Figure 1. Illustration of preference of Pareto loss over Parity loss. In this synthetic data scenario, two subgroups perform at random accuracy level, regardless of the threshold chosen and two other subgroups have higher accuracies when the thresholds are increased. The Pareto loss depicts how far each of the subgroups are from their corresponding optimal accuracy levels. Parity loss depicts the discrepancy between the subgroups accuracies. Parity loss is minimized when all subgroups perform at random accuracy levels, whereas Pareto loss is minimized when all subgroups achieve their optimal accuracies and hence is a better alternative.

Pareto curves is still unknown, a significant result from (Giagkiozis & Fleming, 2012) is presented below, if the Pareto curve is known to be continuous.

If the Pareto-front geometry is continuous, where $f_1, f_2, ..f_M$ denote the M objectives to be optimized which can be parameterized as $f_1^{p_1} + f_2^{p_2} + .. + f_M^{p_M} = C$, such that $p_i > 0$, for a constant C, then for the choice of $p \geq max(p_1, p_2, ...p_M)$, the same guarantees of discoverability from the Tchebycheff problem will hold when using the scalarization, $min((\sum_{m=1}^{M} w_m |f_m(x) - z_m^*|^p)^{\frac{1}{p}})$

In real datasets, the number of points observed on the Pareto front is finite, and hence we usually make the assumption that the Pareto curve is extrapolated using the observed points. Under this assumption, the above lemma would hold on the continuous extrapolated Pareto curve. Also, in our case where we optimize subgroup performance, we know that each $f_i \in [0, 1]$, if accuracy (error) or any other performance metric is scaled. With this tight bound on the performance values, the condition to be satisfied, $f_1^{p_1} + f_2^{p_2} + .. + f_M^{p_M} = C$, becomes trivial to be satisfied empirically under the constraints of numerical precision. For all $p_i = \epsilon$, such that $\epsilon \to 0^+$, we have that $f_i^{p_i} \to 1$ for $f_i(x) \in [0, 1]$. This is evident as $x^\epsilon \to 1$, for $x > 0$ and $lim_{x \to 0^+} x^x = 1$. Since all boundary conditions are also bounded by the limit of 1, we can safely assume $C = M$ and satisfy the condition for most practical purposes, as illustrated in Figure 2. Thus, for choice of $p > \epsilon$, i.e p = 1,2,3.., we see that our weighted metric method produces all discoverable points on the Pareto curve and hence we can be fairly guaranteed (under the errors of numerical precision) that the minimization procedure would find a point on the Pareto curve. Note that the weights of the weighted metric method in our case is based on the fairness criterion and hence all set to 1. This will further impose the fairness constraints during the discovery of points on the Pareto curve.

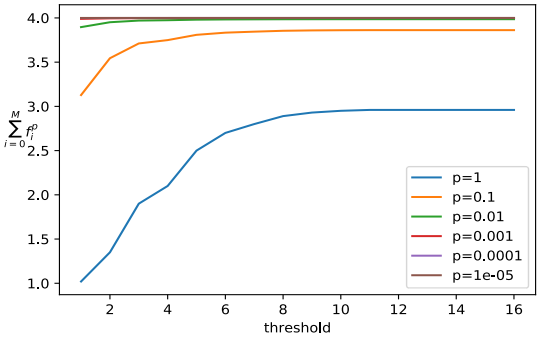

Figure 2. Pareto geometry condition is satisfied for discoverability of all Pareto-Optimal points for Pareto-Efficient Fairness for low values of p, under the errors tolerated by numerical precision

1.3. Efficiency

In this subsection, we provide an analysis of when we expect PEF to outperform standard notions of fairness, like equality of opportunity, i.e when PEF has higher efficiency. (Menon & Williamson, 2018) defines the fairness-frontier which intuitively measures the trade-off between utility (u) (accuracy) and fairness (τ) in the distribution inherent to the problem, rather than one owing to the specific technique one uses, no matter how sophisticated it may be, by computing the fundamental limits of what accuracy is achievable by any classifier. Specifically, the frontier is computed using cost-sensitive measure which quantifies the alignment between the Bayes-optimal plug-in classifier thresholds for the outcome and sensitive attribute distributions. [Proposition 8 in (Menon & Williamson, 2018)]. As the absolute gradient of the fairness frontier increases near the desired fairness constraint, the efficiency gained from using PEF is monotonically non-decreasing.

Specifically, if the fairness frontier shows that for small concessions of the fairness requirement ($\Delta\tau$), the limit of accuracy achievable is much higher (Δu), then we have a possibility that PEF would outperform by choosing such a point on the fairness frontier as shown in Figure **??**. How-

ever, this is a necessary but not sufficient condition. PEF would choose such a point only if the new point is Pareto-dominating the subgroup accuracies of the operating point without the $\Delta\tau$ fairness concession. The amount of concession in Pareto-dominance that we are willing to allow is domain-dependent and can be controlled by tuning the parameter λ in the PEF loss function. Hence, PEF would perform better in conditions where the fairness frontier is steep around the fairness requirement and potential increase in accuracies are achievable in a Pareto-dominant manner.

References

Giagkiozis, I. and Fleming, P. Methods for many-objective optimization: an analysis. 11 2012.

Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 107–118, New York, NY, USA, 23–24 Feb 2018. PMLR.

Miettinen, K. Nonlinear multiobjective optimization, 1999.

Tseng, P. and Yun, S. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1):387–423, Mar 2009. ISSN 1436-4646. doi: 10.1007/s10107-007-0170-0. URL https://doi.org/10.1007/s10107-007-0170-0.

Vincent, M. and Hansen, N. R. Sparse group lasso and high dimensional multinomial classification. *Comput. Stat. Data Anal.*, 71(C):771–786, March 2014. ISSN 0167-9473. doi: 10.1016/j.csda.2013.06.004. URL http://dx.doi.org/10.1016/j.csda.2013.06.004.