arXiv:2307.05728v1 [cs.LG] 11 Jul 2023

Towards A Scalable Solution for Improving Multi-Group Fairness in Compositional Classification

James Atwood [1] **Tina Tian** [1] **Ben Packer** [1] **Meghana Deodhar** [1] **Jilin Chen** [1] **Alex Beutel** [2] **Flavien Prost** [1] **Ahmad Beirami** [1]

Abstract

Despite the rich literature on machine learning fairness, relatively little attention has been paid to remediating complex systems, where the final prediction is the combination of multiple classifiers and where multiple groups are present. In this paper, we first show that natural baseline approaches for improving equal opportunity fairness scale linearly with the product of the number of remediated groups and the number of remediated prediction labels, rendering them impractical. We then introduce two simple techniques, called *task-overconditioning* and *group-interleaving*, to achieve a constant scaling in this multi-group multi-label setup. Our experimental results in academic and real-world environments demonstrate the effectiveness of our proposal at mitigation within this environment.

1. Introduction

The literature around group fairness is relatively rich when we consider a binary classifier and desire to satisfy group fairness for a (binary) group (Dwork et al., 2012; Kamiran et al., 2010; Zafar et al., 2017; Beutel et al., 2017; Agarwal et al., 2018; Donini et al., 2018; Mary et al., 2019; Prost et al., 2019; Baharlouei et al., 2020b; Cho et al., 2020a; Lowy et al., 2022). However, many real-world applications go beyond a single binary decision and we are often faced with multi-label systems where the end decision is a composition of the individual labels (Dwork & Ilvento, 2018; Adomavicius & Tuzhilin, 2005; Burke, 2002; He et al., 2014; Wang et al., 2011).

In this paper, we study a multi-label classification system where several binary classification decisions are combined to make a final prediction for any given input. We consider a special composite classifier where the overall system decision is 1 if any of the individual binary classifier outputs are 1. For example, consider a content moderation system for an online forum that predicts whether a given comment is toxic, insulting, or attacking identity, and hides a comment if any of the predictions are positive (Pavlopoulos et al., 2020). *How do we perform classification based on the combination of these individual predictions, and achieve specific group fairness goals?* This problem, which is a special instance of the compositional fairness (Dwork & Ilvento, 2018; Wang et al., 2021), is the focus of this paper. In addition, we also study a multi-group setting where we are interested in the fairness of the classifier system with respect to many groups

Among the in-process mitigation techniques (Ristanoski et al., 2013; Quadrianto & Sharmanska, 2017; Kamiran et al., 2010; Raff et al., 2018; Aghaei et al., 2019; Donini et al., 2018; Fish et al., 2015; Grari et al., 2020; Cho et al., 2020b; Prost et al., 2019; Lowy et al., 2022; Zafar et al., 2017; Berk et al., 2017; Taskesen et al., 2020; Chzhen & Schreuder, 2020; Baharlouei et al., 2020a; Jiang et al., 2020; Grari et al., 2019), we focus our mitigation strategy on the MinDiff technique (Beutel et al., 2019b; Prost et al., 2019) for improving equality of opportunity (Hardt et al., 2016) in classifiers. MinDiff has proven to be effective at inducing equality of opportunity while maintaining overall classifier performance across a variety of tasks by relying on the maximum mean discrepancy (MMD) estimators (Gretton et al., 2012; Prost et al., 2019). Importantly, MinDiff can be successfully applied to environments when instances labeled with group membership are very sparse, by using a dedicated data streams to ensure that each mini-batch contains a constant number of group labeled examples.

A natural baseline in this scenario is an extension of MinDiff to fairness mitigation in multigroup, multilabel environments, where one regularizer is introduced per each group and per each classifier. However, this baseline causes the batch size to scale linearly in both the number of groups and number of prediction tasks being remediated. Even for small number of groups and and small number of classifiers, this can quickly grow out of hand to the extent that this baseline becomes impractical, especially when the baseline

[1]Google [2]OpenAI (work done while at Google). Correspondence to: James Atwood <atwoodj@google.com>, Tina Tian <ttian@google.com>, Ahmad Beirami <beirami@google.com>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

classifier is already expensive to train. This *significantly* increases resource usage and slows training as the number of groups and prediction tasks grows. We propose two simple optimization techniques to achieve this fairness goal with a constant scaling, with empirical verification. Our contributions are summarized below:

- *Task-overconditioning:* The natural extension of MinDiff requires a batch of negative examples for each label, resulting in a constant scaling with the number of classifiers. Instead, *task-overconditioning* suggests using a single batch that contains the negative examples across all labels. We argue that task-overconditioning further aligns the overall optimization objective to that of mitigating the overall compositional decision, which is our goal, while also achieving a constant scaling with the number of individual classifiers.
- *Group-interleaving:* The natural extension of the mitigation solution requires a batch of negative examples with respect to each group at each iteration. Instead, *group-interleaving* makes the optimization objective stochastic with respect to groups at each iteration, allowing a constant scaling with the number of groups.
- *Empirical verification:* We empirically show that our proposed method, which combines overconditioning and group-interleaving results in equal or better Pareto frontiers than baseline methods, with significant training speedup, on two datasets.

Related Work. Methods for improving group fairness can generally be categorized in three main classes: *pre-processing*, *post-processing*, and *in-processing* methods. Pre-processing algorithms (Feldman et al., 2015; Zemel et al., 2013; Calmon et al., 2017) transform the biased data features to a new space in which the labels and sensitive attributes are statistically independent. Post-processing approaches (Hardt et al., 2016; Pleiss et al., 2017; Alghamdi et al., 2022) achieve group fairness properties by altering the final decision of the classifier.

The focus of this paper is on in-processing methods, which introduce constraints/regularizers for improving fairness in training. These methods have empirically shown to produce a more favorable performance/fairness Pareto tradeoff compared to other methods (Lowy et al., 2022). These include (Ristanoski et al., 2013; Quadrianto & Sharmanska, 2017) decision-trees (Kamiran et al., 2010; Raff et al., 2018; Aghaei et al., 2019), support vector machines (Donini et al., 2018), boosting (Fish et al., 2015), neural networks (Grari et al., 2020; Cho et al., 2020b; Prost et al., 2019; Lowy et al., 2022), or (logistic) regression models (Zafar et al., 2017; Berk et al., 2017; Taskesen et al., 2020; Chzhen & Schreuder, 2020; Baharlouei et al., 2020a; Jiang et al., 2020; Grari et al., 2019). See the recent paper by (Hort et al., 2022) for a more comprehensive literature survey. We focus in this paper on the MinDiff technique, which has been successful across tasks (Beutel et al., 2019b; Prost et al., 2019; Beutel et al., 2019a).

This paper is also broadly related to the compositional fairness literature (Dwork & Ilvento, 2018; Dwork et al., 2020; Wang et al., 2021). In contrast to these works, we focus on a narrower sense of compositionality (only the intersection) for which we derive a scalable specialized solution.

2. Background & Problem Setup

Here, we formally provide the problem setup. Let $(x, \{y_t\}_{t\in[T]})$ represent a feature and a set of T binary labels, where $x \in \mathcal{X}$, and $y_t \in \{0, 1\}$ for all $t \in [T]$[1]. In our setup, the *overall decision* is a simple composite function of the individual labels: $y = \max_{t\in T}\{y_t\}$, i.e., $y = 1$ if and only if there exists $t \in [T]$ s.t. $y_t = 1$.

We consider a scenario where we train T individual predictors $\{\widehat{y}_t(x;\theta)\}_{t\in[T]}$ in a multi-label setup, where $\widehat{y}_t(x;\theta) \in \{0, 1\}$ is a binary classifier from features x, and θ represents model parameters.[2] Similarly, the *overall model prediction* is given by $\widehat{y} = \max_{t\in T}\{\widehat{y}_t\}$, i.e., $\widehat{y} = 1$ if and only if there exists $t \in [T]$ s.t. $\widehat{y}_t = 1$. In other words, we predict that the overall label is 1 when any of the underlying classifiers is triggered. As explained before, this setup is common in many applications, where the final decision (e.g. rejecting a comment (Dixon et al., 2018)) based on $\widehat{y}$ depends on many sub-decisions $\widehat{y}_t$ (properties of the customer or comment).

Our goal is to optimize for fairness in the equal opportunity sense (Hardt et al., 2016) for the overall model prediction with respect to multiple group memberships.[3] Let the set $\mathcal{G}$ capture all groups for which we would like to improve fairness. Let $g_m \in \{0, 1\}$ denote the identifier for membership in group g_m, for $m \in [|\mathcal{G}|]$.

Definition 2.1 (overall equal opportunity with respect to membership in group m)**.**

$$P(\widehat{Y} = 1|G_m = 0, Y = 0) = P(\widehat{Y} = 1|G_m = 1, Y = 0), \tag{1}$$

Note that in this paper, we do not consider the intersectional fairness setting (Kearns et al., 2018; Foulds et al., 2020) where the goal is to ensure fairness to all intersections of group memberships; see Appendix A.

While there are numerous ways to optimize for fairness in machine learning, specifically in equal opportunity sense, the methods that achieve better fairness/performance Pareto frontiers have been empirically observed to be mostly in-processing methods (Lowy et al., 2022), where a regularizer

[1] We define [T]:= $\{1, \ldots, T\}$.

[2] We often drop $(x;\theta)$ for brevity and refer to the output of the t-th classifier as $\widehat{y}_t$.

[3] Fairness of individual predictors is desirable but not required.

is added to the (cross-entropy) training loss to mitigate the model fairness gap. The regularizer is usually of the form: $D(\widehat{Y}(\theta), G_m|Y=0)$ where $D(\cdot,\cdot)$ is a proper divergence between two random variables.

Notice that in our problem setup $\widehat{Y}(\theta)$ is not a differentiable function of the task-level predictors. Hence, we cannot use it directly to regularize the training of the individual task-level classifiers via backpropagation. This situation occurs when task-level predictors are not trained jointly or are even owned by different teams in an organization.

One intuitive solution to remediate this multi-label setup is to ensure that each individual classifier is fair for each group (Dwork & Ilvento, 2018). This intuitive design is motivated by previous work (Wang et al., 2021) which finds that fairness of individual predictors might be sufficient to improve fairness of the overall system, even if there are no theoretical guarantees. We refer to this objective as task-level equal opportunity:

Definition 2.2 (Task-level equal opportunity with respect to group G_m)**.** For any task $t \in [T]$ and group m for $m \in |\mathcal{G}|$,

$$P(\widehat{Y}_t = 1|G_m = 0, Y_t = 0) = P(\widehat{Y}_t = 1|G_m = 1, Y_t = 0). \tag{2}$$

3. Baseline: Many MinDiff Regularizers

While there are many effective methods for solving task-level equal opportunity in (2), as discussed in related work, here we focus on an adaptation of MinDiff (Zafar et al., 2017; Beutel et al., 2019b; Prost et al., 2019) to the multi-group multi-label classification case. This regularization-based approach has a number of advantages. First, it does not require group labels at inference time, which is often true for real-world applications. Next, it has been empirically demonstrated to be effective at remediating fairness issues while still maintaining overall performance (Prost et al., 2019). Finally, it is designed to be effective when group-labeled instances are rare even in training data.

This MinDiff technique introduces a new loss term based on maximum mean discrepancy (MMD) to promote (conditional) independence between the predictions and sensitive group (Prost et al., 2019) per each group and each task. More precisely, the loss becomes:

$$L_{\textit{MinDiff}} = L_{CE}(\hat{Y}, Y) + \lambda \sum_{t \in T} \sum_{m \in [|\mathcal{G}|]} R_{t,m}, \tag{3}$$

where L_{CE} is the empirical cross-entropy loss, λ is a hyper-parameter that sets the relative strength of the entropy and MMD loss,[4] and

$$R_{t,m} = \textit{MMD}(\hat{Y}_t|Y_t=0, G_m=0; \hat{Y}_t|Y_t=0, G_m=1). \tag{4}$$

Computing $R_{t,m}$ requires negative labeled instances ($Y_t = 0$) for both group membership cases ($G_m = 0$ and $G_m = 1$). In practice, instances with group membership information are much less frequently available than those without. Min Diff handles this by creating dedicated data streams for group-labeled instances that ensure that every batch has the data required to compute the MMD kernel component of the loss. As the number of groups and predictions tasks increases, this leads to $O(T \cdot |\mathcal{G}|)$ data streams that must be stored and a $T \cdot |\mathcal{G}|$ multiplier on the batch size.

4. Proposed Method: MinDiff-IO

We now describe our proposed method, *MinDiff-IO*, which is built on two main components: *Group-Interleaving* and *Task-Overconditioning.* The central insight behind these two approaches is that we can still accomplish our goal, overall equal opportunity defined by Equation (1), by optimizing a slightly different objective that is better aligned and is easier to compute. We describe these techniques in the subsequent sections.

4.1. Task-Overconditioning

The baseline method that targets *task-level equal opportunity* has a number of data streams and batch size that scales linearly with T, making it intractable for systems where T might be large (e.g., $O(100)$). Additionally, it does not necessarily imply overall equal opportunity (Dwork & Ilvento, 2018), Equation (1), which is what is desired.

In this section, we present our proposal towards satisfying overall equal opportunity in this compositional decision system. We provide limited theoretical motivation for why it might be more aligned with the overall fairness objective under restrictive assumptions. We shall also see in the experimental section (where those restrictive assumptions are not satisfied) that it leads to equal or better fairness/performance Pareto frontiers.

Definition 4.1 (Overconditioning task-level equal opportunity)**.** For all $t \in [T]$,

$$P(\widehat{Y}_t = 1|G_m = 0, Y = 0) = P(\widehat{Y}_t = 1|G_m = 1, Y = 0). \tag{5}$$

Note that, unlike Equation (4), we condition on all labels having negative truth. This has the effect of requiring only one dataset for all tasks when computing loss rather than T datasets.

Assumption 4.2. Let task-level classifiers $\{\widehat{y}_t(x;\theta)\}_{t \in [T]}$ be such that for all $x \in \mathcal{X}$, and for any $t \neq \tau$,

$$\widehat{y}_t(x;\theta)\widehat{y}_\tau(x;\theta) = 0. \tag{6}$$

[4] In practice, one can tune the MinDiff strength for each regularizer at the expense of a complex hyperparameter tuning.

In other words, the classifiers don't trigger simultaneously; if $\widehat{y}_t = 1$ then $\widehat{y}_\tau = 0$ for all $\tau \neq t$.

Notice that Assumption 4.2 is a strong assumption as it requires the classifiers to have non-overlapping coverage, which is not necessarily satisfied in practice. For example, in the content moderation example, a comment might be toxic, insulting, and attacking identity at the same time. While this assumption is very restrictive, we show that under this scenario overconditioning is perfectly aligned with the goal of mitigating overall classifier. We also don't need this assumption for our empirical results, which show improvements over the baseline classifier.

Lemma 4.3. *If Assumption 4.2 is satisfied, then Definition 4.1 (overconditioned task-level equal opportunity) implies Definition 2.1 (overall equal opportunity).*

The proof is relegated to the appendix. Lemma 4.3 determines a scenario where overconditioning task-level equal opportunity indeed implies the desired overall equal opportunity. Notice that even under Assumption 4.2, Definition 4.1 is a stronger requirement than Definition 2.1, and is not implied by it. In other words, we might be able to satisfy the overall equal opportunity and yet the overconditioning equal opportunity might not be satisfied for all task-level classifiers.

To solve for Task-Overconditioning, we adapt MinDiff loss as follows:

$$L_{\textit{MinDiff-O}} = L_{CE}(\hat{Y}, Y) + \lambda \sum_{t \in T} \sum_{m \in [|\mathcal{G}|]} R^O_{t,m}, \quad (7)$$

where

$$R^O_{t,m} = \textit{MMD}(\hat{Y}_t | Y = 0, G_m = 0; \hat{Y}_t | Y = 0, G_m = 1). \quad (8)$$

Note that there will be fewer data instances that are suitable for computing (8), which requires all labels to be jointly negative, than (4), which only requires individual labels to be negative. We have not found this to be an issue in practical applications where positive label incidence is low.

4.2. Group-Interleaving

MinDiff was originally designed to present remediation data from all groups to the model at each iteration. However, we can reduce the complexity of computing the MinDiff regularizer further by presenting only one group per batch to the model. In this case, the loss becomes:

$$L_{\textit{MinDiff-IO}} = L_{CE}(\hat{Y}, Y) + R^{\mathrm{O}}_M \quad (9)$$

where M is a random index supported on $[|\mathcal{G}|]$. In other words, here we remediate against a random draw from the groups at each iteration of the algorithm. Notice that the new loss is the same as the task-overconditioned loss in expectation, and is expected to converge to a stationary point of the same objective. On the other hand, when combined with Task-Overconditioning, the loss can be computed with only $O(1)$ extra instances in each batch, with no dependence on $|\mathcal{G}|$ and T.

5. Evaluation Metrics

For each binary group membership, i.e., $G_m \in \{0, 1\}$, where $G_m = 1$ is considered the minority group membership, we quantify the fairness gap through the following interchangeable metrics that are expressed in terms of the absolute gap and the ratio of the two groups:

$$d_{EO,m} = |\textit{FPR}_{G_m=1} - \textit{FPR}_{G_m=0}|, \quad (10)$$

and

$$r_{EO,m} = \textit{FPR}_{G_m=1} / \textit{FPR}_{G_m=0}, \quad (11)$$

where $\widehat{P}$ denote the empirical distribution over a test set of N i.i.d samples from P_{XY}, and for $i \in \{0, 1\}$, $\textit{FPR}_{G_m=i} := \widehat{P}(\widehat{Y} = 1 | G_m = i, Y = 0)$.

To measure the classification performance we both compute the Area Under the ROC Curve (ROC AUC) of the classifier as well as accuracy. Finally, to measure speed, we report the number of iterations per second achieved during model training.

6. Experiments

We run two experiments. The first experiment provides the Pareto frontier of fairness vs performance for each approach using a publicly-available academic dataset, and the second provides the performance, fairness, and speed of a real-world policy enforcement classifier at a particular operating point with each of the proposed approaches. Overall, these experiments show that MinDiffIO provides equal or better fairness/performance while improving training speed.

6.1. Civil Comments

The first set of experiments are run on the Civil Comments Dataset (Pavlopoulos et al., 2020); details are given in Appendix B.3. Civil Comments contains comment text and seven associated crowd annotated labels related to the 'civility' of the comment; whether the comment is an insult, toxic, or attacking identity, and so on. We use the subset of the data that are labeled with group information. Groups are related to race, ethnicity, gender, disability, and sexuality.

We train comment classifiers on three of the seven labels and combine the predictions into a system-level prediction: a comment is classified as unsafe is any of the prediction is unsafe. We compare this with 'direct remediation' where

Method	Avg Steps / Sec (↑)	Avg AUCPR (↑)	$r_{EO,1}$ (↓)	$r_{EO,2}$ (↓)
No remediation	A	B	C	D
MinDiff	–	–	–	–
MinDiff-O	$0.60 \times A$	$0.96 \times B$	$0.66 \times C$	$0.79 \times D$
MinDiff-IO	$0.86 \times A$	$0.96 \times B$	$0.65 \times C$	$0.83 \times D$

Table 1. Comparison of different remediation techniques for the policy classification model. Avg AUCPR is a summary measurement of the area under the precision-recall curve for each policy classifier; and Group 1 and Group 2 FPR ratio ($r_{EO,1}$ and $r_{EO,2}$, respectively) denote the ratio of minority to baseline false positive rates at the system level. The baseline MinDiff remediation is too slow to run at this scale (second row). We show how the introduction of Task Overconditioning allows us to remediate at all (third row) and how adding Group Interleaving reduces the speed cost incurred by remediation (fourth row).

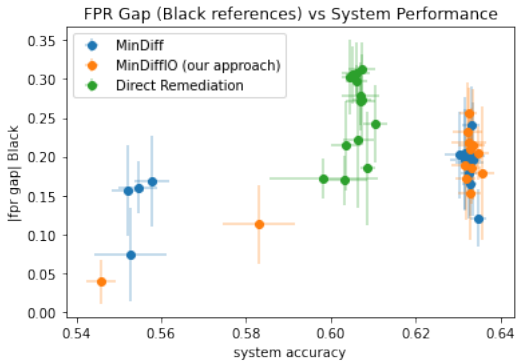

Figure 1. System-level tradeoff curve for the civil comments dataset, achieved by varying regularizer strength λ. Note that direct remediation is outperformed by component-based approaches; for a given level of fairness, the component-based approaches provide better performance. Both component-based approaches offer similar performance. Five runs were performed for each λ and the faint cross-hatches represent 95% confidence intervals.

a classifier is trained to predict the system-level label (the logical OR of the component labels) rather than the components. In addition, we compare with 'component-based' remediation where MinDiff or MinDiffIO are applied to component classifiers. Results are shown for one group (Black) in Figure 1; results for other groups and component-level results for all predictions and groups can be found in Appendix B.3.

The plot displays the tradeoff between fairness and performance as the hyperparameter λ is varied. As λ increases, the contribution of the MMD component of the loss grows, leading to increased fairness at the cost of performance.

We observe that the component-based approaches offer better performance for a given fairness than direct remediation. Also, MindDiff and MinDiffIO offer qualitatively similar results.

6.2. Product Policy Compliance Detection

We now study a real world system, which is responsible for filtering out examples which break the product policy. This is similar to literature on toxic comment detection (Pavlopoulos et al., 2020) or hateful speech filtering (Dixon et al., 2018). To reflect the different facets of the product, a set of rules (10-1000) are defined and an example is against product policy if any given rule is broken. In practice, we use individual classifiers to predict each rule, and an example is filtered out if any individual soft prediction reaches a certain threshold. Samples can be categorized into two sensitive attributes (each considered as binary) and we want to guarantee fairness to samples from each group, which lends itself to the multi-label and multigroup classification.

Note that a false positive of this system is a user harm because policy-following content is flagged as policy-violating. Our goal is to reduce the gaps between false positive rates between minority groups and a baseline population.

We first evaluate the initial system without any remediation and find that two groups have high false positive rate differences. Our goal is to design the mitigation strategy that reduces the observed gaps on the final policy (gap from Definition 2.1) for both groups, while maintaining good performance (measured by AUCPR) and training speed (training steps/sec).

In Table 1, we show four different remediation approaches. The first approach, unremediated, has some performance, fairness, and speed characteristics that we compare other approaches to. The second approach, baseline, is unworkably slow, so we are unable to run experiments or provide results. The third approach, which introduces Task Overconditioning, reduces fairness gaps with a minor hit to performance and a major hit to speed. Finally, the fourth approach adds Group Interleaving to mitigate the speed impact while maintaining similar fairness and performance characteristics.

7. Conclusion

Prior in-process equal opportunity remediation methods suffer from poor (linear) scaling in the number of prediction tasks and number of groups to remediate, making existing techniques sometimes impossible to apply to real-world scenarios. We present Mindiff-IO, a new method that builds on the MinDiff approach to provide constant scaling with respect to tasks and groups. We show that Mindiff-IO provides similar performance and fairness characteristics to MinDiff while scaling much better in multilabel and multigroup en-

vironments through experiments with both academic and real-world datasets. The limitations of this work are provided in Appendix A.

Acknowledgements

We would like to thank Preethi Lahoti, Ananth Balashankar, Lucian Cionca, and Katherine Heller for their constructive feedback on this paper.

References

Adomavicius, G. and Tuzhilin, A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005. doi: 10.1109/TKDE.2005.99.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *International Conference on Machine Learning*, pp. 60–69. PMLR, 2018.

Aghaei, S., Azizi, M. J., and Vayanos, P. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1418–1426, 2019.

Alghamdi, W., Hsu, H., Jeong, H., Wang, H., Michalak, P., Asoodeh, S., and Calmon, F. Beyond adult and compas: Fair multi-class prediction via information projection. *Advances in Neural Information Processing Systems*, 35: 38747–38760, 2022.

Baharlouei, S., Nouiehed, M., Beirami, A., and Razaviyayn, M. Rényi fair inference. In *ICLR*, 2020a.

Baharlouei, S., Nouiehed, M., Beirami, A., and Razaviyayn, M. Rényi fair inference. In *International Conference on Learning Representations*, 2020b.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

Beutel, A., Chen, J., Zhao, Z., and Chi, E. H. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H., et al. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2212–2220, 2019a.

Beutel, A., Chen, J., Doshi, T., Qian, H., Woodruff, A., Luu, C., Kreitmann, P., Bischof, J., and Chi, E. H. Putting fairness principles into practice: Challenges, metrics, and improvements. *CoRR*, abs/1901.04562, 2019b. URL http://arxiv.org/abs/1901.04562.

Burke, R. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12: 331–370, 2002.

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pp. 3992–4001, 2017.

Cho, J., Hwang, G., and Suh, C. A fair classifier using mutual information. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2521–2526. IEEE, 2020a.

Cho, J., Hwang, G., and Suh, C. A fair classifier using kernel density estimation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b.

Chzhen, E. and Schreuder, N. A minimax framework for quantifying risk-fairness trade-off in regression. *arXiv preprint arXiv:2007.14265*, 2020.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pp. 67–73, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278729. URL https://doi.org/10.1145/3278721.3278729.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pp. 2791–2801, 2018.

Dwork, C. and Ilvento, C. Individual fairness under composition. *Proceedings of Fairness, Accountability, Transparency in Machine Learning*, 2018.

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

Dwork, C., Ilvento, C., and Jagadeesan, M. Individual fairness in pipelines. *arXiv preprint arXiv:2004.05167*, 2020.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Fish, B., Kun, J., and Lelkes, A. D. Fair boosting: a case study. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning*. Citeseer, 2015.

Foulds, J. R., Islam, R., Keya, K. N., and Pan, S. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pp. 1918–1921. IEEE, 2020.

Grari, V., Ruf, B., Lamprier, S., and Detyniecki, M. Fairness-aware neural Réyni minimization for continuous features. *arXiv preprint arXiv:1911.04929*, 2019.

Grari, V., Hajouji, O. E., Lamprier, S., and Detyniecki, M. Learning unbiased representations via Rényi minimization. *arXiv preprint arXiv:2009.03183*, 2020.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

He, X., Pan, J., Jin, O., Xu, T., Liu, B., Xu, T., Shi, Y., Atallah, A., Herbrich, R., Bowers, S., et al. Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the eighth international workshop on data mining for online advertising*, pp. 1–9, 2014.

Hort, M., Chen, Z., Zhang, J. M., Sarro, F., and Harman, M. Bia mitigation for machine learning classifiers: A comprehensive survey. *arXiv preprint arXiv:2207.07068*, 2022.

Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, pp. 862–872. PMLR, 2020.

Kamiran, F., Calders, T., and Pechenizkiy, M. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*, pp. 869–874. IEEE, 2010.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pp. 2564–2572, 2018.

Lowy, A., Baharlouei, S., Pavan, R., Razaviyayn, M., and Beirami, A. A stochastic optimization framework for fair risk minimization. *Transactions of Machine Learning Research*, 2022.

Mary, J., Calauzenes, C., and El Karoui, N. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, pp. 4382–4391. PMLR, 2019.

Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., and Androutsopoulos, I. Toxicity detection: Does context really matter?, 2020.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pp. 5680–5689, 2017.

Prost, F., Qian, H., Chen, Q., Chi, E. H., Chen, J., and Beutel, A. Toward a better trade-off between performance and fairness with kernel-based distribution matching. *arXiv preprint arXiv:1910.11779*, 2019.

Quadrianto, N. and Sharmanska, V. Recycling privileged learning and distribution matching for fairness. *Advances in Neural Information Processing Systems*, 30, 2017.

Raff, E., Sylvester, J., and Mills, S. Fair forests: Regularized tree induction to minimize model bias. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 243–250, 2018.

Ristanoski, G., Liu, W., and Bailey, J. Discrimination aware classification for imbalanced datasets. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pp. 1529–1532, 2013.

Taskesen, B., Nguyen, V. A., Kuhn, D., and Blanchet, J. A distributionally robust approach to fair classification. *arXiv preprint arXiv:2007.09530*, 2020.

Wang, L., Lin, J., and Metzler, D. A cascade ranking model for efficient ranked retrieval. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pp. 105–114, 2011.

Wang, X., Thain, N., Sinha, A., Prost, F., Chi, E. H., Chen, J., and Beutel, A. Practical compositional fairness: Understanding fairness in multi-component recommender systems. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 436–444, 2021.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pp. 962–970. PMLR, 2017.

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

A. Limitations

There are three limitations to the approaches mentioned here. First, overconditioning requires instances that have negative ground truth for all modeled labels[5] in order to compute the Min Diff loss. This is a realistic environment; for instance, a policy dataset where policy-violating content is rare. However, if true positives are very common, this method may no longer be effective.

The second limitation is with respect to intersectional group fairness. The interleaving optimization described in Section 4.2 does not explicitly represent or remediate the intersection of groups. Intersectional remediation is a more difficult problem due to the exponential scaling of the number of intersections with respect to the number of groups. We opted not to remediate intersections because of the sparsity of our group labels - very few instances are labeled with more than one group. We believe that techniques that effectively and efficiently address intersectional remediation are an interesting area for future work.

Third, we only consider MinDiff-based techniques in this paper and demonstrate that Mindiff-IO has better scaling characteristics than the original MinDiff approach. Future work could compare the fairness, performance, and scaling properties of Mindiff-IO with other methods of achieving equal opportunity. In addition, future work could test the application of interleaving and overconditioning to other in-processing methods.

B. Proofs, Experiment Details, and Further Results

B.1. Proof of Lemma 4.3

Proof of Lemma 4.3. The proof is completed by noting that

$$\begin{aligned} P(\widehat{Y}=1|G=0,Y=0) &= P(\max_{t\in[T]} \widehat{Y}_t = 1|G=0,Y=0) \\ &= \sum_{t\in[T]} P(\widehat{Y}_t = 1|G=0,Y=0) \qquad (12) \\ &= \sum_{t\in[T]} P(\widehat{Y}_t = 1|G=1,Y=0) \qquad (13) \\ &= P(\max_{t\in[T]} \widehat{Y}_t = 1|G=1,Y=0) \qquad (14) \\ &= P(\widehat{Y}=1|G=1,Y=0), \end{aligned}$$

where (12) follows from Assumption 4.2, and (13) follows from Definition 4.1, and (14) follows from Assumption 4.2. □

B.2. Civil Comments Experimental Details

For these experiments we select three labels (identity attack, insult, and toxicity) as well as four groups (black, gay or lesbian, female, and transgender) for modeling and remediation. Our model consists of a single hidden layer deep neural network that takes a simple hashing trick bag of words vectorization of the comment text as input. The hidden layer and text vector have 64 and 1,000 elements, respectively.

All models are trained for 25 epochs with a learning rate if 0.1 and a Gaussian kernel weight of 1.0.

We present empirical Pareto frontiers of fairness (here, the absolute value of the difference between false positive rates for a group and baseline) and performance (here, ROC AUC). Thresholds for the fairness dimension are selected through calibration on a validation set.

B.3. Detailed Civil Comments Results

System-level results for all four groups are shown in Figure 2. Note that, in each case, component-based techniques outperform direct remediation by offering a higher performance for a given fairness.

Component results for each group, label pair are shown in Figure 3 for both the MinDiff and Mindiff-IO techniques. These

[5]Note that the ground truth negative requirement is present when optimizing for equality of opportunity with respect to false positive rates. If equality of opportunity with respect to false negative rates were the goal, the method would instead require ground truth positives.

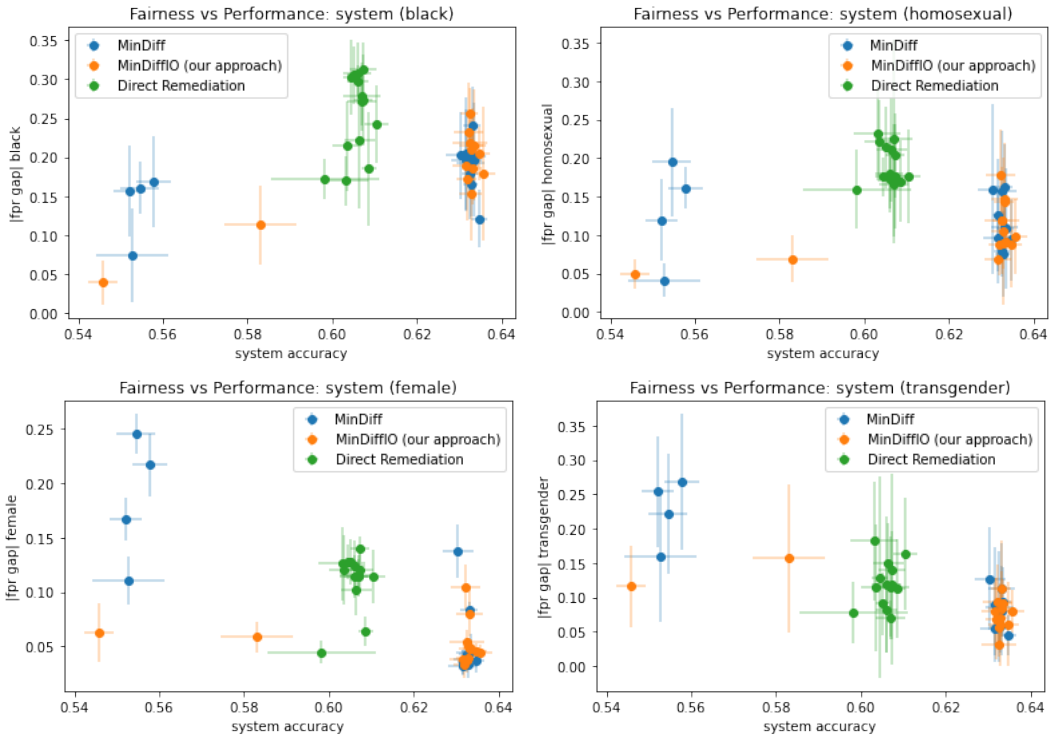

Figure 2. System-level results for all four groups. Note that, in each case, component-based techniques outperform direct remediation by offering a higher performance for a given fairness.

Pareto frontiers are generated by varying the hyperparameter λ, where higher λ values put more weight on the MinDiff loss term and lead to improved fairness at the cost of performance. Each data point in the plot is generated by training a model five times; the crosses in each dimension represent 95% confidence intervals.

Note that each approach achieves a similar Pareto frontier, indicating the Mindiff-IO has similar performance and fairness characteristics. In other words, this experiment confirms that Mindiff-IO does not sacrifice classifier fairness or performance for individual classifiers. In the next experiment, we will provide training speed measurements to demonstrate the scaling advantages of Mindiff-IO.

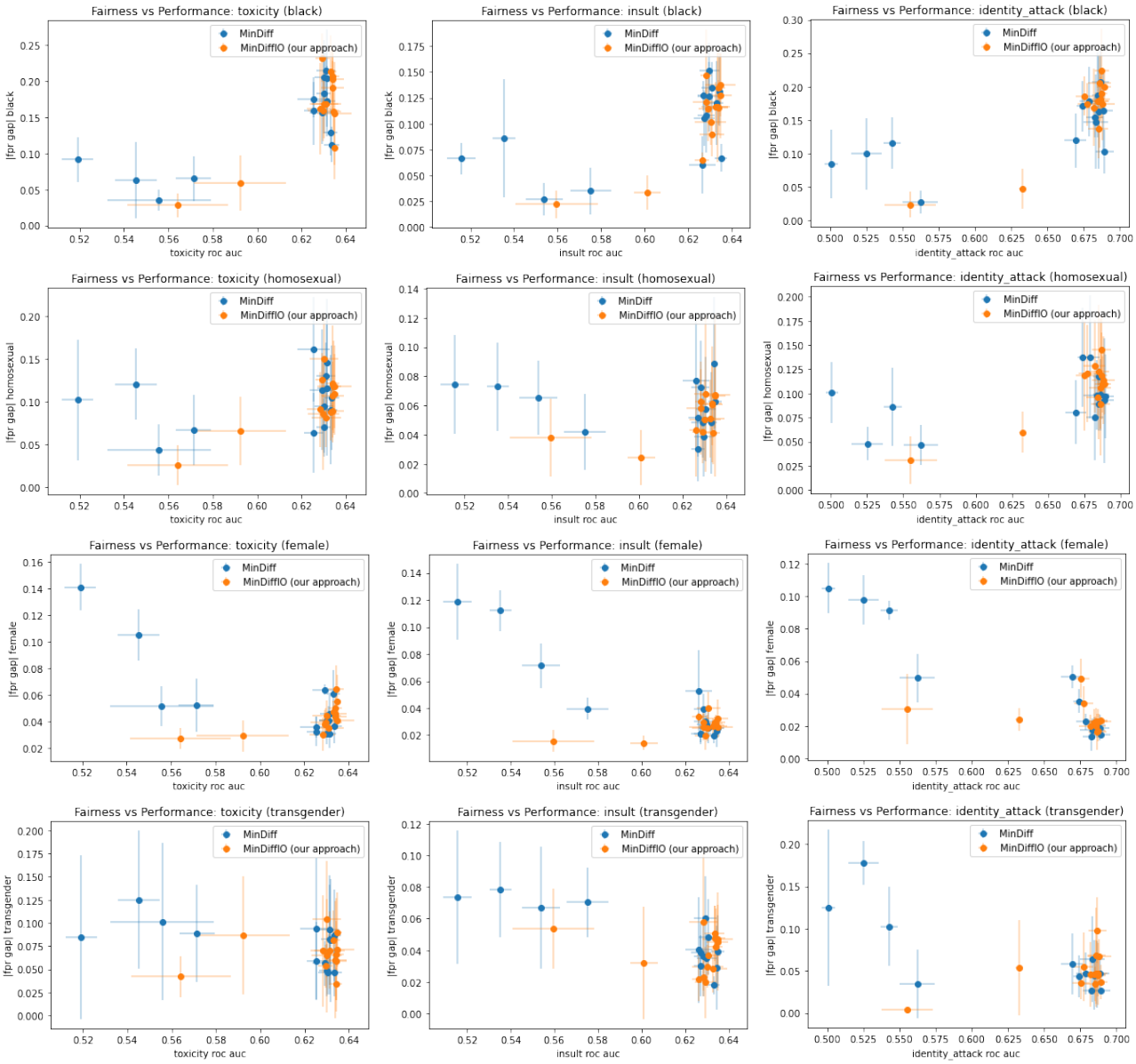

Figure 3. Empirical Pareto frontiers of fairness and performance for baseline MinDiff and newly introduced Mindiff-IO. Each point corresponds to a different min diff strength, and the error bars in each dimension represent 95% confidence intervals. Note that the frontier achieved by Mindiff-IO not qualitatively different than that achieved by the baseline MinDiff approach. Note that there is a clear frontier that trades off performance and fairness for groups with a large FPR gap (black and homosexual). However, for groups with a low FPR gap (female, transgender), a small min diff strength λ provides the best performance and fairness characteristics.