Pairwise Fairness for Ranking and Regression

Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, Serena Wang
Google Research
{hnarasimhan, acotter, mayagupta, serenawang}@google.com

August 14, 2019

Abstract

We present pairwise fairness metrics for ranking and regression models that form analogues of statistical fairness notions such as equal opportunity, equal accuracy, and statistical parity. Our pairwise formulation supports both discrete protected groups, and continuous protected attributes. We show that the resulting training problems can be efficiently and effectively solved using constrained optimization and robust optimization techniques based on two player game algorithms developed for fair classification. Experiments illustrate the broad applicability and trade-offs of these methods.

1 Introduction

As machine-learned ranking and regression models become more prevalent and have a greater impact on people's day-to-day lives, it is urgent that we have better tools to quantify, measure, track, and improve fairness metrics for ranking and regression models. Algorithms for respecting fairness metrics have been relatively well-studied for binary classification and categorical protected groups. Here, we build upon the relatively mature work in quantifying and training fairer classifiers, extending those ideas and algorithms to ranking and regression models.

We draw inspiration from the standard learning-to-rank strategy [23]: reducing the ranking problem to that of learning a binary classifier to predict the relative ordering of *pairs* of examples. Our central observation is that we can express a broad set of statistical fairness metrics in terms of such pairwise comparisons, where the choice of *which pairs are considered* is used to specify different notions of fairness. Given a fairness goal expressed in terms of accuracy on pairs, we will show how one can use state-of-the-art constrained optimization algorithms to effectively train ranking and regression models with improved fairness metrics. This re-framing of the problem into pairwise comparisons also empowers fairer ways to handle protected inputs that are continuous.

Given a set of ranked pairs of results, the key question remains how to define the fairness metric. Similar to the binary classification setting, we believe there is not one right fairness definition, and instead we provide a paradigm that makes it easy to define and train for different fairness definitions, analogous to those useful in classification. A key split is between unsupervised and supervised fairness metrics. For example, consider the task of recommending restaurants to college students who prefer cheaper restaurants, and consider French and Mexican as protected groups. Our unsupervised statistical parity constraint would require that it is equally likely that (i) a French restaurant is ranked above a Mexican restaurant, and (ii) a Mexican restaurant is ranked above a French restaurant. In contrast, our supervised *equal opportunity* constraint would require that its equally likely that (i) a cheap French restaurant is ranked above an expensive Mexican restaurant, and (ii) a cheap Mexican restaurant is ranked above an expensive French restaurant.

The following are the main contributions of this paper:

Table 1: Example of group-dependent pairwise accuracy matrix of $K = 3$ groups, and the frequency-weighted (FW) row average and column average.

Better Result	Worse Result: French	Mexican	Chinese	FW Row Mean
French	*0.76*	0.68	0.56	0.70
Mexican	0.61	*0.62*	0.55	0.60
Chinese	0.62	0.61	*0.56*	0.60
FW Col. Mean	0.69	0.65	0.56	*0.65*

1. We define the notion of a group-dependent pairwise accuracy for a ranking function (Section 2) and use this to construct ranking analogues of (supervised) *equal opportunity* and *equal accuracy* criteria (Section 4). We also provide a pairwise analogue of (unsupervised) *statistical parity* for ranking.

2. We show that proposed pairwise metrics extend easily to regression problems (Section 5) by forming pairs based on the ordering induced by the regression model. We similarly handle continuous protected attributes by forming pairs based on the ordering of the protected feature (Section 6).

3. We extend existing machinery for fair binary classification to optimize for the proposed pairwise fairness metrics. Specifically, we use the state-of-the-art constrained optimization algorithms of Cotter et al. [10] to develop two approaches: one where the pairwise metrics are imposed as constraints, and the other where they are enforced through a robust optimization setup (Section 7).

4. We present extensive experiments on a variety of fair ranking and regression tasks with both discrete and continuous protected attributes (Section 8). Our results illustrate the benefit of directly optimizing for pairwise fairness metrics compared to an existing sampling/weighting heuristic [13].

2 Pairwise Fairness

Suppose that we have a set of K protected groups $G_1, \ldots, G_K$ partitioning the space of examples $\mathcal{X} \times \mathcal{Y}$ (i.e. every example belongs to exactly one group). We define the *group-dependent pairwise accuracy* $A_{G_i > G_j}$ as the accuracy of a ranking function $f : \mathcal{X} \to \mathbb{R}$ on those pairs for which the labeled "better" example is a member of group G_i, and the labeled "worse" example is a member of group G_j. That is:

$$A_{G_i > G_j} := P(f(x) > f(x') \mid y > y', (x, y) \in G_i, (x', y') \in G_j), \tag{1}$$

where (x, y) and (x', y') are drawn *i.i.d.* from the distribution of examples, restricted to the appropriate protected groups. Given K groups, one can compute the $K \times K$ matrix of all possible K^2 group-dependent pairwise accuracies (see Table 1). One may also be interested in how a specific group is doing on average:

$$A_{G_i > :} := P(f(x) > f(x') \mid y > y', (x, y) \in G_i) \tag{2}$$

$$A_{: > G_i} := P(f(x) > f(x') \mid y > y', (x', y') \in G_i). \tag{3}$$

The accuracy in (2) is averaged all pairs for which the G_i example was labeled as "better," and the "worse" example is from any group, including G_i. Similarly, (3) is the accuracy averaged over all pairs where G_i example should not have been prefered. Lastly, the overall pairwise accuracy $P(f(x) > f(x') \mid y > y')$ is simply the standard AUC.

These group-dependent pairwise metrics will enable us to construct ranking and regression analogues of the equalized odds fairness metric [17] and the equal accuracy metric [10]. The correspondence between pairwise comparisons and pairwise binary classification enables us to borrow machinery developed for training binary classifiers with fairness constraints to training ranking and regression models with constraints on the propose pairwise fairness metrics.

3 Related Work

We review related work that we build upon in fair classification, and then related work on the problems addressed here: fair ranking, fair regression, and handling continuous protected attributes. We also point of similarities/differences with recently proposed AUC-based metrics for fairness.

3.1 Fair Classification

Many statistical fairness metrics for binary classification can be written in terms of *rate constraints*, that is, constraints on the classifier's positive (or negative) prediction rate for different groups [16, 25, 9, 10]. For example, the goal of *demographic parity* [15] is to ensure that the classifier's positive prediction rate is the same across all protected groups. Similarly, the *equal opportunity* metric [17] requires that true positive rates, which can be expressed in terms of the classifier's positive prediction rates on positively-labeled examples, should be equal across all protected groups. Many other statistical fairness metrics can be expressed in terms of rates, e.g. *equal accuracy*, *no worse off* and *no lost benefits* [10]. Constraints on these fairness metrics can be added to the training objective for a binary classifier, then solved using constrained optimization algorithms or relaxations thereof [32, 16, 33, 14, 1, 9, 10]. Here, we extend this work to train ranking and regression models with pairwise fairness constraints.

3.2 Fair Ranking

A majority of the previous work on fair ranking has focused on list-wise definitions for fairness that depend on the entire list of results for a given query [e.g. 35, 7, 5, 29, 34, 30]. These include both *unsupervised* criteria that require the average exposure near the top of the ranked list to be equal for different groups [e.g. 29, 7, 34], and *supervised* criteria that require the average exposure for a group to be proportional to the average relevance of that group's results to the query [5, 29, 30]. Of these, some provide post-processing algorithms for re-ranking a given ranking [5, 7, 29, 30], while others learn a ranking model from scratch [34, 30].

There have also been two recent papers on pairwise fairness notions [4, 19], both of which are focused on ranking with categorical groups and are methodologically different from us.

Beutel et al. [4] propose ranking pairwise fairness definitions equivalent to those we give in (1), (2) and (3). Their work focuses on ranking and on categorical groups, whereas we show how to use pairwise formulations to capture a wider variety of different statistical fairness notions, and to handle regression and continuous protected features as well as ranking. Another substantial difference is in how the two papers propose optimizing a model to improve pairwise fairness: they relax the problem to minimizing the correlation between the residual between a clicked and unclicked item and the group membership of the clicked item. In contrast, we give a more direct and flexible approach to train with one's choice of pairwise fairness constraints (see Section 7.5 for more details). Cotter et al. [10] have noted that such constraint formulations are easier for practitioners to specify and test. Previous work comparing relaxing to correlation vs directly solving the constrained optimization problem has shown substantial wins for the more direct approach we take here [16]. Lastly, there are major experimental differences: they provide an in-depth case study of one real-world recommendation problem, whereas we provide a broad set of experiments on public and real-world data illustrating the effectiveness on both ranking and regression problems, for categorical or continuous protected attributes.

Kallus and Zhou [19] also provide pairwise fairness metrics based on AUC for bipartite ranking problems, but only consider categorical groups and propose a post-processing approach that fits a monotonic transform to an existing ranking model to optimize the proposed metrics. In contrast, we additionally handle regression problems and continuous protected attributes, and develop more flexible approaches that directly optimize for the desired pairwise fairness goals during training.

3.3 Fair Regression

Defining fairness metrics for a regression setting is a challenging task, with a long history in standardized testing [18]. Komiyama et al. [21] consider the unfairness of a regressor in terms of the correlation between the output and a protected attribute. Pérez-Suay et al. [27] regularize to minimize the Hilbert-Schmidt independence between the protected features and model output. These definitions have the "flavor" of statistical parity, in that they attempt to remove information about the protected feature from the model's predictions. Here, we focus more on supervised fairness notions like equal opportunity and equal accuracy.

Berk et al. [3] propose regularizing linear regression models for the notion of fairness corresponding to the principle that *similar individuals receive similar outcomes* [15]. Their definitions focus on enforcing similar squared error, which fundamentally differs from our definitions in that we assume each group would prefer higher scores, not necessarily more accurate scores.

More recently, Agarwal et al. [2] propose a *bounded group loss* fairness definition which insists that the regression error be within an allowable limit for each protected group. In contrast, our pairwise equal opportunity definitions for regression do not rely on a specific regression loss, but instead are based on the ordering induced by the regression model within and across groups. Agarwal et al. also propose a statistical parity definition for regression based on histogram matching of the output distributions for different groups.

3.4 Continuous Protected Features

Most prior work in machine learning fairness has assumed categorical protected groups, in some cases extending those tools to continuous features by bucketing [20]. Fine-grained buckets raise statistical significance challenges, and coarse-grained buckets raise unfairness issues of how the lines between bins are drawn and due to the lack of protection within each bin. Raff et al. [28] considered continuous protected features in their tree-growing criterion that addresses fairness. Kearns et al. [20] focused on statistical parity type constraints for continuous protected features for classification. Komiyama et al. [21] controlled the correlation of the model output with protected variables (which may be continuous). Mary et al. [24] propose a fairness criterion for continuous attributes based on the Rényi maximum correlation coefficient. *Counterfactual fairness* [22, 26] requires that changing a protected attribute, while holding causally unrelated attributes constant, should not change the model output distribution, but this does not directly address issues with ranking fairness.

3.5 Pinned AUC

The matrix representation of the group-dependent pairwise accuracy metrics defined in Section 1 and illustrated in Table 1 has a strong relationship to the pinned AUC fairness metric introduced by Dixon et al. [13]. With two protected groups, pinned AUC works by resampling the data such that each of the two groups make up 50%, and then calculating the ROC AUC on the resampled dataset. Based on the well-known equivalence between ROC AUC and average pairwise accuracy, Borkan et al. [6] demonstrate that pinned AUC, as well as their new proposed metric, weighted pinned AUC, can be decomposed as a linear combination of within-group and cross-group pairwise accuracies. In other words, both pinned AUC and weighted pinned AUC can be written as linear combinations of the different pairwise accuracies $A_{G_i > G_j}$.

Thus pinned AUC is one possible weighting of the pairwise accuracy matrix. We propose a broader set of fairness criteria for ranking and regression that are expressed as constraints on entries of the pairwise accuracy matrix. Our provided optimization algorithms automatically find the right weighting of the matrix entries to satisfy the specified fairness criteria (see Section 7.3 for details).

In our experiments, we compare our algorithms against (a version of) the de-biasing sampling-based approach of Dixon et al. [13]. This serves as a representative baseline that optimizes a fixed weighting of the pairwise accuracy matrix.

4 Ranking Pairwise Fairness Metrics

We consider a standard ranking set-up [23]: we're given a sample S of queries drawn *i.i.d.* from an underlying distribution $\mathcal{D}$, where each query is a set of candidates to be ranked, and each candidate is represented by an associated feature vector $x \in \mathcal{X}$ and label $y \in \mathcal{Y}$. The label space can be, for example, $\mathcal{Y} = \{0, 1\}$ (e.g. for click data: $y = 1$ if a result was clicked by a user, $y = 0$ otherwise), $\mathcal{Y} = \mathbb{R}$ (each result has an associated quality rating), or $\mathcal{Y} = \mathbb{N}$ (the labels are a ground truth ranking). We adopt the convention that higher labels should be ranked closer to the top. Any of these choices of label space $\mathcal{Y}$ support forming *pairs* of candidates $((x, y), (x', y'))$ belonging to the same query, where $y > y'$ as in Section 2.

For ranking problems, we use the pairwise accuracy definitions of Section 2 with one additional constraint on the choice of pairs: we never form pairs from two examples belonging to two different queries. Cross-query comparisons can be okay, but require labels that are comparable across ranked lists. Next, we use the pairwise accuracies to define *supervised* pairwise fairness goals and *unsupervised* fairness notions.

4.1 Pairwise Equal Opportunity

We construct a *pairwise equal opportunity* analogue of the *equal opportunity* metric [17]:

$$A_{G_i > G_j} = \kappa, \text{ for some } \kappa \in [0, 1], \text{ for all } i, j \tag{4}$$

Equal opportunity for binary classifiers [17] requires positively-labeled examples to be equally likely to be predicted positively regardless of protected group membership. Similarly, this *pairwise equal opportunity* for ranking problems requires *pairs* to be equally-likely to be ranked correctly regardless of the protected group membership of both members of the pair. By symmetry, we could equally well consider $A_{G_i > G_j}$ to be a true positive rate or a true negative rate, so there is no distinction between "equal opportunity" and "equal odds" in the ranking setting, when all of the pairwise accuracies are constrained equivalently.

Pairwise equal opportunity can be relaxed to be less constraining, either by (i) requiring all pairwise accuracies to be above some desired lower bound (e.g. $A_{G_i > G_j} \geq 0.7$ for all $i \neq j$), (ii) to be within some quantity of each other (e.g. $\max_{i \neq j} A_{G_i > G_j} - \min_{i \neq j} A_{G_i > G_j} \leq 0.1$), or (iii) requiring the worst group-dependent accuracy $A_{G_i > G_j}$ to be as large as possible (i.e. $\text{maximize } \min_{i \neq j} A_{G_i > G_j}$) in the style of robust optimization [e.g. 8]. We will show in Section 7 how models can be efficiently trained subject to each of these different types of pairwise fairness constraints.

4.2 Within-Group vs. Cross-Group Comparison

We have observed that labels for within-group comparisons ($i = j$) are sometimes more accurate and consistent across raters, whereas the labels for cross-group comparisons ($i \neq j$) can be noisier and less consistent. For example, consider a video ranking system where group i is sports and group j is cooking shows. Experts in sports could be more likely to rate sports videos, and to rate them accurately, but on average may not provide as useful ratings if asked to compare a sports video to a cooking video.

For these reasons, one may wish to *separately* constrain within-group comparisons and cross-group comparisons. To that end, we define two requirements. The first, which we call *cross-group pairwise equal opportunity*, only constrains cross-group comparisons:

$$A_{G_i>G_j} = \kappa, \text{ for some } \kappa \in [0,1], \text{ for all } i \neq j \tag{5}$$

The second constrains for *within-group pairwise equal accuracy*:

$$A_{G_i>G_i} = \kappa', \text{ for some } \kappa' \in [0,1], \text{ for all } i. \tag{6}$$

Separating these metrics enables one to can constrain them independently, with possibly different slack. In certain applications, particularly those in which cross-group comparisons are rare or do not occur, we might want to constrain *only* pairwise equal accuracy (6). For example, we might want a music ranking system to be equally accurate at ranking jazz as it is at ranking country music, to be fair to their constituent audiences, but if the two types of music appear relatively rarely on the same ranked list it may not be statistically worthwhile to try to also constrain the cross-group ranking accuracy (5).

4.3 Marginal Equal Opportunity

The previous pairwise equal opportunity proposals are defined in terms of the K^2 group-dependent pairwise accuracies. This may be too fine-grained, either for statistical significance reasons, or because the fine-grained constraints might be infeasible. To address this, we propose a looser *marginal pairwise equal opportunity* criterion that asks for parity for ech group averaged over the other groups:

$$A_{G_i>:} = \kappa \text{ for some } \kappa \in [0,1], \text{ for } i = 1, \dots, K. \tag{7}$$

4.4 Statistical Parity

Finally, our pairwise setup can also be used to define unsupervised fairness metric. A *pairwise statistical parity* constraint can be defined as:

$$P\left(f(x) > f(x') \mid (x,y) \in G_i, (x',y') \in G_j\right) = \kappa \text{ for any } i \neq j. \tag{8}$$

This constraint requires that if two candidates are compared from different groups, then on average each group has an equal chance of being top-ranked. This constraint completely ignores the training labels, but that may be useful when groups are so different that any comparison is too *apples-to-oranges*, or raters are not expert enough to make useful cross-group comparisons.

5 Regression Pairwise Fairness Metrics

The proposed pairwise fairness definitions extend elegantly to the standard regression setting, in which $f : \mathcal{X} \to \mathcal{Y}$ attempts to predict a regression label for each example. This is not a ranking problem, so there are no queries—instead, we have a training set of N examples, and we compute pairwise metrics over the all N^2 pairs (or a random subset, if N^2 is too large).

5.1 Regression Equal Opportunity

Using these pairs in the regression setting, one can compute and constrain the pairwise equal opportunity metrics as in (4), (5), (6) and (7) for regression models. For example, restricting (5) constrains the model to be equally likely for all groups G_i and G_j to assign a higher score to group i examples over group j examples, if the group i example's label is higher. Note this fairness notion is asymmetric: we treat higher scores as more (or less) desirable, and we seek to control how often each group gets high scores. This asymmetric perspective is applicable if the scores confer a benefit, such as regression models that estimate credit scores or admission to college, or if the model scores dictate a penalty to be avoided, such as getting stopped by police. This asymmetry assumption that getting higher scores is either preferred or not-preferred is analogous to the binary classification case where a positive label is assumed to confer some benefit.

5.2 Regression Symmetric Cross-Group Equal Accuracy

For regression problems where each group's goal is to be accurate (rather than to score high or low), one can define symmetric pairwise fairness metrics as well, for example, the symmetric pair accuracy for group as G_i is $A_{G_i>:} + A_{:>G_i}$, and one might constrain these accuracies to be the same across groups. This strategy is different than trying to equate the squared loss across groups, whereas this metric is purely ordinal and related to the Kendall-Tau metric [23].

5.3 Regression Equal Accuracy

Promoting *pairwise equal accuracy* as per (6) for regression asks that the model be equally faithful for every group to the pairwise ranking of any two within-group examples. This is especially useful if the regression labels from different groups arise from different communities that have different labeling distributions. For example, suppose all jazz music examples are rated by jazz lovers who only give 4-5 star ratings, but all classical music examples are rated by critics who give a range of 1-5 star ratings, with 5 being rare. Simply training to minimize MSE might be dominated by the classical music score examples, because the classical errors are likely to be larger and affect the MSE more.

5.4 Regression Statistical Parity

The pairwise statistical parity condition described in (8) can also be applied to the regression setting. It requires, "Comparing any two examples from two different groups, they are equally likely to have the higher score." One sufficient condition to guarantee pairwise statistical parity is if the distribution of outputs $f(X)$ for a random input X is the same for each of the protected groups. This condition can be enforced approximately by histogram matching the output distributions for different protected groups [2].

6 Continuous Protected Features

Suppose we have a continuous or ordered protected feature Z, for example, we may wish to constrain for fairness with respect to age, income, seniority, etc. The proposed pairwise fairness notions extend nicely to this setting by changing how the pairs are formed to depend on the ordering of the protected feature, rather on the protected groups. Specifically, we change (1) to the following *continuous attribute pairwise accuracies*:

$$A_{>} := P(f(x) > f(x') \mid y > y', z > z'), \tag{9}$$
$$A_{<} := P(f(x) > f(x') \mid y > y', z < z'), \tag{10}$$

where z is the protected feature value for (x, y) and z' is the protected feature value for (x', y'). For example, if the protected feature Z measures *height*, then $A_{>}$ measures the accuracy of the model when comparing pairs where the candidate who is taller should receive a higher score.

The previously proposed pairwise fairness constraints for discrete groups have analogous definitions in this setting by replacing (1) with (9). Pairwise equal opportunity becomes

$$A_{>} = A_{<}, \tag{11}$$

which requires, for example, that the model be equally accurate when the taller or shorter candidate should be top-ranked.

Similar to Section 4, $A_<$ can be thought of as true negative rate for pairs $(x, y), (x', y')$ where $z > z'$, and by symmetry, $A_<$ is also equal to the true positive rate for pairs where $z < z'$:

$$\begin{aligned} A_< &= P(f(x) < f(x') \mid y < y', z > z') \\ &= P(f(x) > f(x') \mid y > y', z < z') \\ &= TPR_{z<z'}. \end{aligned}$$

Therefore, (11) equates both the true positive rates and the true negative rates for both sets of pairs, and specifies both equalized odds and equal opportunity.

7 Training for Pairwise Fairness

First, we show how one can use the pairwise fairness definitions to specify a training objective for ranking or regression or continuous protected features with fairness constraints. Then we show how to optimize these objectives by borrowing existing machinery for fair binary classification. For ease of exposition, we will first describe our approaches for the case of the ranking and cross-group equal opportunity criteria, and then explain how to adapt these optimization problems to other settings and goals.

7.1 Proposed Optimization Problem Formulation

Let $A(f)$ denote the group-dependent pairwise accuracy matrix whose i-jth entry $A_{G_i>G_j}(f)$ is defined by (1) for a ranking model $f : \mathcal{X} \to \mathbb{R}$. Let $AUC(f)$ be the overall pairwise accuracy (equivalently the area under the ROC curve). Let $\mathcal{F}$ be the class of models we are interested in. We consider formulating fairness goals using constrained optimization with an allowed slack ϵ, or as a robust optimization problem:

$$\textbf{Constrained opt.:} \max_{f\in\mathcal{F}} AUC(f) \text{ s.t. } A_{G_i>G_j}(f) - A_{G_k>G_l}(f) \le \epsilon \;\; \forall i \neq j, k \neq l. \tag{12}$$

$$\textbf{Robust opt.:} \max_{f\in\mathcal{F},\xi} \xi \text{ s.t. } \xi \le AUC(f), \xi \le A_{G_i>G_j}(f) \;\; \forall i \neq j. \tag{13}$$

These two problem formulations can be varied to handle the other pairwise fairness metrics we have proposed. For regression problems, we use the constrained optimization formulation with the maximum AUC objective replaced with minimum mean squared error as the objective.

7.2 Reduction to Rate Constraints

Both the constrained and robust optimization formulations can be written in terms of binary *rate constraints* on score differences [16, 10]. Re-write a pairwise accuracy term as,

$$A_{G_i>G_j}(f) = \mathbb{E}\left[\mathbf{I}_{f(x)-f(x')>0} \,\middle|\, ((x,y),(x',y')) \in \mathcal{S}_{ij}\right],$$

where $\mathbf{I}$ is the usual indicator function and $\mathcal{S}_{ij} = \{((x,y),(x',y')) \in (\mathcal{X}\times\mathcal{Y})^2 \,|\, y > y', (x,y) \in G_i, (x',y') \in G_j\}$. This enables us to adopt the proxy-Lagrangian framework of Cotter et al. [9, 10] to solve the optimization problems in (12) and (13). Similar to Agarwal et al. [1] and Kearns et al. [20], this framework learns a *stochastic model* that is supported on a finite set of functions in $\mathcal{F}$. The high-level idea is to set up a two-player game, where one player optimizes the model parameters, and the other player optimizes over the space of Lagrange multipliers (with the rate constraints replaced with differentiable proxy constraints for the first player). The result of solving the two-player game is a stochastic model that satisfies the rate constraints in expectation.

7.3 Connection to Weighted Pairwise Accuracy

As discussed in Section 3, pinned AUC is a previous fairness metrics that are weighted sums of the entries of the pairwise accuracy matrix: $\sum_{i,j} \beta_{i,j} A_{G_i > G_j}(f)$. The constrained and robust optimization formulations that we propose are equivalent to maximizing the weighted pairwise accuracy for a specific, optimal set of weights $\{\beta_{i,j}\}$ (see Kearns et al. [20], Agarwal et al. [1] for relationship between constrained optimization and cost-weighted learning). The constrained optimization solver essentially finds the optimal set of weights $\{\beta_{i,j}\}$ for the two formulations. This is convenient because it enables us to specify fairness criteria in terms of pairwise accuracies, rather than having to manually tune the $\{\beta_{i,j}\}$ needed to produce the pairwise accuracies we want.

7.4 Extension to Other Set-ups

The formulations described in Section 7.1 easily apply to other pairwise fairness goals and to regression problems. To apply the formulations to the marginal equal opportunity criterion, we would define similar constraints on row-based averages $A_{G_i >:}(f)$ for all i (see (2)). To apply the formulations to continuous protected features, we would define constraints on the protected pairwise accuracies $A_{>}(f)$ and $A_{<}(f)$ (see (9)). For regression problems, we adopt the constrained optimization formulation with the objective replaced with the (negative) mean squared error. We do not use the robust optimization approach for regression tasks as the squared error is not necessarily comparable with the regression pairwise metrics.

7.5 Relationship with Beutel et al. [4]

While the ranking fairness metrics of Beutel et al. [4] are equivalent to those we give in (1), (2) and (3), there are two significant differences between their *overall approach*, and ours. First, they add a fixed fairness penalty to their training problem, which has the advantage of being parameter-free, but the disadvantage of granting the user no ability to control the fairness vs. accuracy trade-off. Our approach *(i)* enables the user to choose a point on the Pareto front via their choice of fairness constraint, and *(ii)* enables us to leverage existing work on constrained optimization for machine learning, including the state-of-the-art algorithms like [1, 8], and the provably improved-generalization strategy of Cotter et al. [11].

Second, in Beutel et al., there are only two protected groups, whereas we enable the user to constrain any number of groups, with the constrained-optimization algorithm automatically determining how much each of the groups must be penalized/rewarded in order to satisfy the constraints. If you extend the fixed regularizer approach of Beutel et al. to multiple groups there are no such parameters to trade-off, so it'd be possible for one protected class to "dominate" the others. One could fix this by introducing separate penalties for each protected group, and then trying to choose their values appropriately, but this is essentially what our approach does *automatically*.

8 Experiments

We illustrate our proposals for improving pairwise fairness metrics on five ranking problems and two regression problems. We implemented the proposed constrained and robust optimization methods described in Section 7 in TensorFlow using the *proxy Lagrangian* optimization in the open-source TensorFlow Constrained Optimization toolbox [9, 10]. Our code will be made publicly available.

8.1 Experimental Details

We implemented the constrained and robust optimization methods described in Section 7 in TensorFlow using the open-source *proxy Lagrangian* optimization toolbox [9, 10]. We use Adam for gradient updates, and

Table 2: Test AUC with test pairwise fairness violations in parentheses. For fairness violations, we report $|A_{G_0>G_1} - A_{G_1>G_0}|$ when imposing cross-group constraints, $\max\{|A_{G_0>G_1} - A_{G_1>G_0}|,\ |A_{G_0>G_0} - A_{G_1>G_1}|\}$ when imposing both cross-group and in-group constraints, $|A_{G_0>:} - A_{G_1>:}|$for marginal constraints, and $|A_{>} - A_{<}|$ for the constraint on continuous protected attributes.

Dataset	Prot. Group	Unconstrained	Debiased	Constrained	Robust
Sim. CG	0/1	**0.919** (0.275)	**0.919** (0.275)	0.864 (**0.008**)	0.857 (0.018)
Sim. CG and IG	0/1	**0.919** (0.275)	**0.919** (0.275)	0.750 (0.054)	0.888 (**0.053**)
Business Matching	Business Type	**0.695** (0.062)	0.698 (0.061)	0.685 (**0.003**)	0.685 (0.071)
Wiki Talk Page	Term "Gay"	**0.974** (0.104)	0.972 (0.014)	0.962 (**0.010**)	0.941 (0.025)
W3C Experts	Gender	0.528 (0.963)	0.539 (0.895)	**0.544** (**0.100**)	0.542 (0.143)
Crime	Race Percent	**0.931** (0.184)	–	0.810 (**0.036**)	0.858 (**0.036**)

use hinge loss based proxy constraints. For the Wiki Talk Pages dataset and the Law School dataset, we use minibatches of 100 stochastic gradients to better handle the large number of pairs to be enumerated. In each case, the dataset is split into training, validation and test sets in the ratio 1/2:1/4:1/4, with the validation set used to tune the learning rate, the number of training epochs for the unconstrained optimization methods, and for the post-processing shrinking step in the proxy-Lagrangian solver. For datasets with queries, we evaluate the pairwise accuracy metrics for individual queries and report the average across queries. The constrained and robust optimization methods learn a stochastic model, and the metrics reported for these methods are expectations over random draw of a scoring function.

8.2 Pairwise Fairness for Ranking

We experimented with five different ranking problems.

8.2.1 Debiasing Comparison

For our ranking experiments, we also compare against an algorithm that attempts to imitate the *debiasing* scheme of Dixon et al. [13] by optimizing a weighted pairwise accuracy (without any explicit constraints):

$$\frac{1}{n_+ n_-} \sum_{i,j: y_i > y_j} \alpha_{z_i, y_i} \alpha_{y_j, z_j} \mathbf{1}(f(x_i) > f(x_j)),$$

where n_+, n_- are the number of positively labeled and negatively labeled training examples, and $\alpha_{z,y} > 0$ is a non-negative weight on each label and protected group are set such that the relative proportions of positive and negative examples within each protected group are balanced. Specifically, we fix $\alpha_{0,-1} = \alpha_{0,+1} = \alpha_{1,+1} = 1$ and set $\alpha_{1,-1}$ so that $\frac{|\{x_i \,|\, z_i=0, y_i=-1\}|}{|\{x_i \,|\, z_i=0, y_i=+1\}|} = \alpha_{1,-1} \frac{|\{x_i \,|\, z_i=1, y_i=-1\}|}{|\{x_i \,|\, z_i=1, y_i=+1\}|}$. This mimics Dixon et al. [13] where they sample additional negative documents belonging group 1, so that the relative label proportions within each group are similar.

8.2.2 Simulated Ranking Data

For this toy ranking task with two features, there are 10,000 queries, and each query has 11 candidates. For each query, we randomly pick one of the 11 candidates to have a positive label $y = +1$ and the other 10 candidates receive a negative label $y = -1$, and we independently randomly assign each candidate's protected attribute z *i.i.d.* from a $Bernoulli(0.1)$ distribution. Then we generate two features simulated to score how well the candidate matches the query, from a Gaussian distribution $\mathcal{N}(\mu_{y,z}, \Sigma_{y,z})$, where

	Negative	
Positive	**Group 0**	**Group 1**
Group 0	0.941	0.980
Group 1	0.705	0.894

(a) Unconstrained

	Negative	
Positive	**Group 0**	**Group 1**
Group 0	0.854	**0.900**
Group 1	**0.907**	0.927

(b) Constrained/Cross-group

	Negative	
Positive	**Group 0**	**Group 1**
Group 0	*0.743*	**0.769**
Group 1	**0.779**	*0.797*

(c) Constrained/All

	Negative	
Positive	**Group 0**	**Group 1**
Group 0	***0.881***	0.930
Group 1	0.897	0.935

(d) Robust/All

Figure 1: Test pairwise accuracy matrix for simulated (ranking) data.

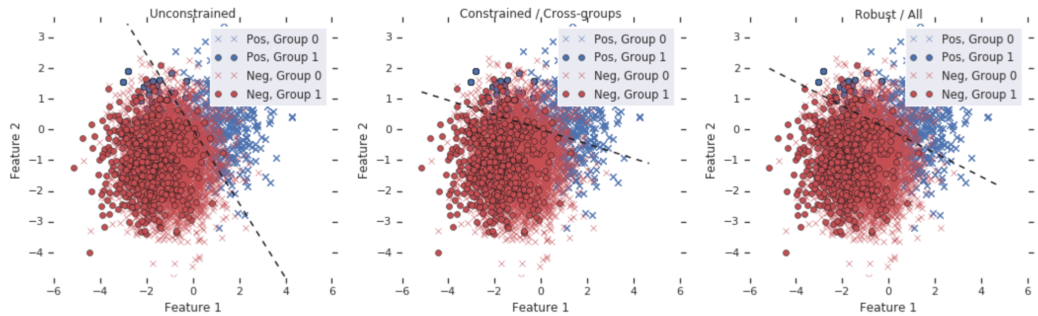

Figure 2: Plot of learned hyperplanes on simulated ranking data. For constrained and robust optimization, we plot the hyperplane that is assigned the highest probability in the support of the learned stochastic model.

$\mu_{-1,0} = \begin{bmatrix}-1\\-1\end{bmatrix}$, $\mu_{-1,1} = \begin{bmatrix}-2\\-1\end{bmatrix}$, $\mu_{+1,0} = \begin{bmatrix}1\\0\end{bmatrix}$, $\mu_{+1,1} = \begin{bmatrix}-1.5\\0.75\end{bmatrix}$ and $\Sigma_{-1,0} = \Sigma_{-1,1} = \Sigma_{+1,0} = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$, $\Sigma_{+1,1} = \begin{bmatrix}0.5 & 0\\0 & 0.5\end{bmatrix}$.

We train linear ranking functions $f(x) : \mathbb{R}^2 \to \mathbb{R}$. For the 1st experiment, we seek to enforce *cross-group equal opportunity* with constrained optimization by constraining $|A_{0>1} - A_{1>0}| \leq 0.01$). For the robust optimization, we implement this goal by maximizing $\min\{A_{0>1}, A_{1>0}, AUC\}$. For the 2nd experiment we seek to enforce *cross-group equal opportunity and in-group equal accuracy* by constraining both $|A_{0>1} - A_{1>0}| \leq 0.01$ and $|A_{0>0} - A_{1>1}| \leq 0.01$. For the robust optimization, we implement this goal by maximizing $\min\{A_{0>1}, A_{1>0}\}$ + $\min\{A_{0>0}, A_{1>1}\}$.

Table 2 gives the test ranking accuracy and pairwise fairness goal violations. Debiasing is not helpful in this case. Only the constrained optimization achieves the first fairness goal.

For the 2nd experiment, the robust optimization does better with the second fairness goal. Figure 1 shows the 2×2 pairwise accuracy matrix for each method. From Figure 1(c) one sees that constrained optimization satisfies the fairness constraints by lowering the accuracies for all four group-pairs. In contrast, Figure 1(d) shows robust optimization maximizes the minimum entry in the fairness matrix. These results are consistent with the two different optimization problems: you get what you ask for.

Figure 2 shows the hyperplanes (dashed line) for the different linear ranking functions learned by the different methods. The hyperplane learned by the unconstrained approach ranks the majority examples (the + group) well, but is not accurate at ranking the minority examples (the o group). The hyperplanes learned by the constrained and robust optimization methods work more equally well for both groups.

Note that the quality of the learned ranking function depends on the slope of the hyperplane and is unaffected by its intercept.

The pairwise fairness violations are measured as $|A_{0>1} - A_{1>0}|$ for the first fairness criterion and as

	Chain	Non-chain
Row-avg	0.767	0.706

(a) Unconstrained

	Chain	Non-chain
Row-avg	0.780	0.719

(b) Debiased

	Chain	Non-chain
Row-avg	**0.715**	**0.712**

(c) Constrained/Row-avg

	Chain	Non-chain
Row-avg	0.804	0.733

(d) Robust/Row-avg

Figure 3: Test row-based matrix averages on business match (ranking) data.

$\max\{|A_{0>1} - A_{1>0}|, |A_{0>0} - A_{1>1}\}$ for the second criterion.

As expected the unconstrained algorithm yields the highest overall ranking objective, but incurs very high fairness violations. The debiased weighting approach also gives a similar performance (as the relative proportions of positives and negatives are the same in expectation for the two protected groups since the protected group was independent of the label). Both the constrained and robust optimization approaches significantly reduce the fairness violations. In terms of the ranking objective, constrained optimization suffers a considerable decrease in the overall objective when constraining all entries of the accuracy matrix, whereas robust optimization incurs only a marginal decrease in objective.

Whereas constrained optimization performed better on differences in the matrix entries, robust optimization maximizes the smallest entry.

8.2.3 Business Matching

This is a proprietary dataset from a large internet services company of ranked pairs of relevant and irrelevant businesses for different queries, with a total of 17,069 pairs. How well each query matches each candidate is represented by 41 features. We consider two protected groups, *chain* and *not chain*. We define a candidate as a member of the *chain* group G_{chain} if its query is seeking a chain business and the candidate is a chain business. We define a candidate as a member of the *non-chain* group $G_{non-chain}$ if its query is not seeking a chain business and the candidate is not a chain business. Candidates do not belong to either of these two groups if they are chain and the query is not-chain-seeking, or vice-versa.

For this experiment, we experiment with imposing a marginal equal opportunity constraint: $|A_{chain>:} - A_{non-chain>:}| \leq 0.01$. This requires the model to be roughly as accurate at correctly matching chain businesses to chain-seeking queries as it is at correctly matching non-chain businesses to non-chain-seeking queries. With robust optimization, we maximize $\min\{A_{chain>:}, A_{non-chain>:}, AUC\}$. All comparisons trained a two-layer neural network model with 10 hidden nodes. As seen in Table 2, compared to the unconstrained approach, the constrained optimization yields very low fairness violation, while only being marginally worse on the test AUC. Taking a closer look at each group's pairwise accuracies (see Figure 3), we find that robust optimization yields the best matrix row marginals for both groups - that is, robust optimization was the most accurate for the two groups, but its overall accuracy was poor because of poorer behavior on the examples that were not covered by these two groups. Debiasing produced a negligible reduction in fairness violation, but yields better row marginals than the unconstrained approach.

8.2.4 Wiki Talk Page Comments

This public dataset contains 127,820 comments from Wikipedia Talk Pages labeled with whether or not they are toxic (i.e. contain "rude, disrespectful or unreasonable" content [13]). This is a dataset where debiased weighting has been effective in learning fair, unbiased classification models [13]. We consider the task of learning a ranking function that ranks comments that are labeled toxic higher than the comments that are labeled non-toxic, in order to help the model's users identify toxic comments. Following Dixon et al. [13], we

Toxic \ Non-toxic	Other	Gay
Other	0.973	0.882
Gay	0.986	0.937

(a) Unconstrained

Toxic \ Non-toxic	Other	Gay
Other	0.971	**0.953**
Gay	**0.967**	0.945

(b) Debiased

Toxic \ Non-toxic	Other	Gay
Other	0.962	**0.943**
Gay	**0.953**	0.922

(c) Constrained/Cross-groups

Toxic \ Non-toxic	Other	Gay
Other	0.940	**0.920**
Gay	**0.945**	0.927

(d) Robust/Cross-groups

Figure 4: Test pairwise accuracy matrix on Wiki Talk Page comments (ranking) data.

Table 3: Regression test MSE (lower is better) and test pairwise fairness violation (in parenthesis).

Dataset	Prot. Group	Constraint Type	Unconstrained	Constrained
Law School	*Gender*	Cross-group	**0.142** (0.304)	0.143 (**0.019**)
Crime	*Race Percentage*	Continuous Attr.	**0.021** (0.327)	0.028 (**0.032**)

consider the protected attribute defined by whether the term 'gay' appears in the comment. Among comments that have the term 'gay', 55% are labelled toxic, whereas among comments that do not have the term 'gay', only 9% are labelled toxic. We learn a convolutional neural network model with the same architecture used in Dixon et al. [13]. We consider a cross-group equal opportunity criterion. We impose $|A_{Other>Gay} - A_{Gay>Other}| \leq 0.01$ with constrained optimization and maximize $\min\{A_{Other>Gay},\, A_{Gay>Other}, AUC\}$ with robust optimization. As seen in Table 2, by re-weighting the pairs to have the same relative label proportions within each group, the debiasing approach reduces the fairness violation considerably. The constrained optimization approach yields even lower fairness violation, but at the cost of a slightly lower test AUC. The group pairwise accuracies are shown in Figure 4. Among the cross-group errors, the learned model is more likely to incorrectly rank a non-toxic comment with the term 'gay' higher than a toxic comment without the term.

8.2.5 W3C Experts Search

We also evaluate our methods on the W3C Experts dataset, previously used to study disparate exposure in ranking [34]. This is a subset of the TREC 2005 enterprise track data, and consists of 48 topics and 200 candidates per topic, with each candidate labeled as an expert or non-expert for the topic. The task is to rank the candidates based on their expertise on a topic, using a corpus of mailing lists from the World Wide Web Consortium (W3C). This is an application where the unconstrained algorithm does better for the minority protected group. We use the same features as Zehlike and Castillo [34] to represent how well each topic matches each candidate; this includes a set of five aggregate features derived from word counts and tf-idf scores, and the gender protected attribute.

For this task, we learn a linear model and impose a cross-group equal opportunity constraint: $|A_{Female>Male} - A_{Male>Female}| \leq 0.01$. For robust optimization, we maximize $\min\{A_{Female>Male},\, A_{Male>Female},\, AUC\}$. As seen in Table 2 and the group pairwise accuracies are shown in Figure 5 , the unconstrained ranking model incurs a huge fairness violation. This is because the unconstrained model treats gender as a strong signal of expertise, and often ranks female candidates over male candidates. As a result, $A_{Female>Male}$ is close to 100%, while $A_{Male>Female}$ is close to 0. Not only do the constrained and robust optimization methods achieve significantly lower fairness violations, they also happen to produce higher test metrics due to the constraints acting as regularizers and reducing overfitting.

		Non-expert	
		Male	**Female**
Expert	**Male**	0.534	0.028
	Female	0.991	0.573

(a) Unconstrained

		Non-expert	
		Male	**Female**
Expert	**Male**	0.541	0.081
	Female	0.977	0.571

(b) Debiased

		Non-expert	
		Male	**Female**
Expert	**Male**	0.540	**0.501**
	Female	**0.601**	0.571

(c) Constrained/Cross-groups

		Non-expert	
		Male	**Female**
Expert	Male	0.540	**0.471**
	Female	**0.614**	0.553

(d) Robust/Cross-groups

Figure 5: Test pairwise accuracy matrix for W3C experts (ranking) data.

		Low	
		Male	**Female**
High	Male	0.653	0.490
	Female	0.795	0.654

(a) Unconstrained

		Low	
		Male	**Female**
High	**Male**	0.652	**0.647**
	Female	**0.666**	0.655

(b) Constrained/Cross-groups

Figure 6: Pairwise accuracy matrix for law school (regression).

8.2.6 Communities and Crime

We next illustrate handling a *continuous protected attribute* in a ranking problem. We use the benchmark *Communities and Crime* dataset from UCI [12] which contains 1,994 communities in the United States described by 140 features, and the per capita crime rate for each community. As in prior work [9], we label the communities with a crime rate above the 70th percentile as 'high crime' and the others as 'low crime', and consider the task of learning a ranking function that ranks high crime communities above the low crime communities. We treat the percentage of black population in a community as a continuous protected attribute.

We learn a linear ranking function, with the protected attribute included as a feature. We do not compare to debiasing as it does not apply to continuous protected attributes. Adopting the continuous attribute equal opportunity criterion in Section 6, we impose the constraint $|A_< - A_>| \leq 0.01$. Table 2 shows the constrained and robust optimization methods reduce the fairness violation by more than half, at the cost of a slightly lower test AUC.

8.3 Pairwise Fairness for Regression

We present experiments for two regression problems; results in Table 3.

Law School: This dataset [31] contains details of 27,234 law school students, and we predict the undergraduate GPA for a student from the student's LSAT score, family income, full-time status, race, gender and the law school cluster the student belongs to, with gender as the protected attribute. We impose a cross-group equal opportunity constraint: $|A_{Female>Male} - A_{Male>Female}| \leq 0.01$. Table 3 shows the constrained optimization approach successfully massively reduces the fairness violation compared to the unconstrained model, at only a smalll increase in squared error. The group pairwise accuracies are the Law School regression problem are shown in Figure 6.

Communities and Crime: This dataset has continuous labels for the per capita crime rate for a community. Once again, we treat the percentage of black population in a community as a *continuous protected attribute* and impose a continuous attribute equal opportunity constraint: $|A_> - A_<| \leq 0.01$. The results are shown in Table 3 and Figure 7. The constrained approach yields a huge reduction in fairness violation, though at the cost of an increase in squared error.

$A_>$	$A_<$
0.904	0.577

(a) Unconstrained

$A_>$	$A_<$
0.777	**0.745**

(b) Constrained/Continuous Attr.

Figure 7: Continuous attribute pairwise accuracies for crime (regression) with percentage of black population as protected attribute.

9 Conclusions and Further Ideas

We showed that supervised and unsupervised notions of fairness for ranking, regression, and handling continuous protected features can be intuitively specified using pairwise metrics; and that doing so enables leveraging state-of-the-art constrained optimization and robust optimization solvers. We compared to the recent de-biasing heuristic [13], and were able to reproduce the original paper's good results on Wiki Talk [13], but did not find the heuristic performed well for other problems.

The key way one specifies the proposed pairwise fairness metrics is by the selection of which pairs to consider. Here, we focused on within-group and cross-group pairs. One could also bring in side information such as position in ranking, or for regression, weight example pairs based on their label differences.

References

[1] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. M. Wallach. A reductions approach to fair classification. In *ICML*, 2018.

[2] A. Agarwal, M. Dudik, and Z. S. Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *ICML*, 2019.

[3] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. A convex framework for fair regression. *FATML*, 2017.

[4] A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi, and C. Goodrow. Fairness in recommendation through pairwise experiments. *KDD Applied Data Science Track*, 2019. URL `arxiv.org/abs/1903.00780.pdf`. [To appear].

[5] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *ACM SIGIR*, 2018.

[6] D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman. Robust metrics for measuring unintended bias with real data for text classification, 2019.

[7] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. Ranking with fairness constraints. In *ICALP*, 2018.

[8] R. S. Chen, B. Lucier, Y. Singer, and V. Syrgkanis. Robust optimization for non-convex objectives. In *NIPS*, 2017.

[9] A. Cotter, H. Jiang, and K. Sridharan. Two-player games for efficient non-convex constrained optimization. In *ALT*, 2019. URL `https://arxiv.org/abs/1804.06500`. [To appear].

[10] A. Cotter, H. Jiang, S. Wang, T. Narayan, M. R. Gupta, S. You, and K. Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *JMLR*, 2019. URL `arxiv.org/pdf/1809.04198.pdf`. [To appear].

[11] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In *ICML*, 2019.

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

[13] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. Measuring and mitigating unintended bias in text classification. In *AIES*, 2018.

[14] M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints, 2018.

[15] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *ITCS*, 2012.

[16] G. Goh, A. Cotter, M. R. Gupta, and M. P. Friedlander. Satisfying real-world goals with dataset constraints. In *NIPS*, pages 2415–2423. 2016.

[17] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *NIPS*, 2016.

[18] J. E. Hunter and F. L. Schmidt. Critical analysis of the statistical and ethical implications of various definitions of test bias. *Psychological Bulletin*, 1976.

[19] Nathan Kallus and Angela Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. *arXiv preprint arXiv:1902.05826*, 2019.

[20] M. Kearns, S. Neel, A. Roth, and S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *ICML*, 2018.

[21] J. Komiyama, A. Takeda, J. Honda, and H. Shimao. Nonconvex optimization for regression with fairness constraints. In *ICML*, 2018.

[22] M. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. 2017.

[23] T.-Y. Liu. *Learning to Rank for Information Retrieval*. Springer, 2011.

[24] J. Mary, C. Calauzènes, and N. E. Karoui. Fairness-aware learning for continuous attributes and treatments. In *ICML*, 2019.

[25] H. Narasimhan. Learning with complex loss functions and constraints. In *AIStats*, 2018.

[26] J. Pearl, M. Glymour, and N. Jewell. *Causal Inference in Statistics: a Primer.* Wiley, 2016.

[27] A. Pérez-Suay, V. Laparra, G. Mateo-García, J. Muñoz-Marí, L. Gómez-Chova, and G. Camps-Valls. Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2017.

[28] E. Raff, J. Sylvester, and S. Mills. Fair forests: Regularized tree induction to minimize model bias. *AAAI/ACM Conference on Artificial Intelligence, Ethics, Society*, 2018.

[29] A. Singh and T. Joachims. Fairness of exposure in rankings. In *KDD*, 2018.

[30] Ashudeep Singh and Thorsten Joachims. Policy learning for fairness in ranking. *arXiv preprint arXiv:1902.04056*, 2019.

[31] L. Wightman. LSAC national longitudinal bar passage study. *Law School Admission Council*, 1998.

[32] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: A mechanism for fair classification. In *ICML Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2015.

[33] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *AIStats*, 2017.

[34] M. Zehlike and C. Castillo. Reducing disparate exposure in ranking: A learning to rank approach, 2018. URL `https://arxiv.org/abs/1805.0871`.

[35] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. A. Baeza-Yates. Fa*ir: A fair top-k ranking algorithm. In *CIKM*, 2017.