Pairwise Fairness for Ranking and Regression

Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, Serena Wang
Google Research
1600 Amphitheatre Pkwy, Mountain View, CA 94043
{hnarasimhan, acotter, mayagupta, serenawang}@google.com

Abstract

We present pairwise fairness metrics for ranking models 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 existing constrained optimization and robust optimization techniques developed for fair classification. Experiments illustrate the broad applicability and trade-offs of these methods.

Introduction

As ranking models and regression models become more prevalent and have a greater impact on people's day-to-day lives, it is important that we develop better tools to quantify, measure, track, and improve fairness metrics for such models. A key question for ranking and regression is how to define fairness metrics. As in the binary classification setting, we believe there is not one "right" fairness definition: instead, we provide a paradigm that makes it easy to define and train for different fairness definitions, analogous to those that are popular for binary classification problems.

One key distinction is between unsupervised and supervised fairness metrics: for example, consider the task of ranking restaurants for college students who prefer cheaper restaurants, and suppose we wish to be fair to French vs Mexican restaurants. Our proposed unsupervised *statistical parity* constraint would require that the model be equally likely to (i) rank a French restaurant above a Mexican restaurant, and (ii) rank a Mexican restaurant above a French restaurant. In contrast, our proposed supervised *equal opportunity* constraint would require that the model be equally likely to (i) rank a cheap French restaurant above an expensive Mexican restaurant, and (ii) rank a cheap Mexican restaurant above an expensive French restaurant.

Like some recent work on fair ranking (Beutel et al. 2019; Kallus and Zhou 2019), we draw inspiration from the standard learning-to-rank strategy (Liu 2011): we reduce the ranking problem to that of learning a binary classifier to predict the relative ordering of *pairs* of examples. This reduction of ranking to binary classification enables us to formulate a broad set of statistical fairness metrics, inspired by analogues in the binary classification setting, in terms of pairwise comparisons. The same general idea can actually be applied more broadly: in addition to group-based fairness in the ranking setting, we show that the same overall approach can also be applied to (i) the regression setting, or (ii) the use of continuous protected attributes instead of discrete groups. In all three of these cases, we show how to effectively train ranking models or regression models to satisfy the proposed fairness metrics, by applying state-of-the-art constrained optimization algorithms.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Ranking Pairwise Fairness Metrics

We begin by considering a standard ranking set-up (Liu 2011): 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 $\mathcal{Y}$ induce a partial ordering on examples, for which all candidates belonging to the *same* query are totally ordered, and any two candidates (x, y) and (x', y') belonging to *different* queries are incomparable.

Suppose that we have a set of K protected groups $G_1, \ldots, G_K$ partitioning the space of examples $\mathcal{X} \times \mathcal{Y}$ such that 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 belongs to group G_i, and the labeled "worse" example belongs to 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. Notice that this definition implicitly forces us to construct pairs only from examples belonging to the same query, since y and y' are not comparable if they belong to different queries—however, the probability is taken over all such pairs, across *all* queries. Given K groups, one can compute the $K \times K$ matrix of all possible K^2 group-dependent pairwise accuracies. One can also measure how each group performs *on average*:

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

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

The accuracy in (2) is averaged over 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 the G_i example should not have been preferred. Lastly, the overall pairwise accuracy $P(f(x) > f(x') \mid y > y')$ is simply the standard AUC. Next, we use the pairwise accuracies to define *supervised* pairwise fairness goals and *unsupervised* fairness notions.

Pairwise Equal Opportunity

We construct a *pairwise equal opportunity* analogue of the *equal opportunity* metric (Hardt, Price, and Srebro 2016):

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

Equal opportunity for binary classifiers (Hardt, Price, and Srebro 2016) 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 either by requiring all pairwise accuracies (i) to only 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 (ii) only requiring the minimum pairwise accuracy $A_{G_i>G_j}$ to be as big as possible (i.e. $\text{maximize } \min_{i \neq j} A_{G_i>G_j}$), in the style of robust optimization [e.g. Chen et al. 2017]. We will later show how models can be efficiently trained subject to both these types of pairwise fairness constraints using existing algorithms.

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 than labels for cross-group comparisons ($i \neq j$) can be noisier and less consistent. This especially arises when the labels are coming from experts that are more comfortable with rating candidates from certain groups. For example, consider a video ranking system where group i is sports videos and group j is cooking shows. If our experts can choose which videos they rate (as in most consumer recommendation systems with feedback), sports experts are likely to rate sports videos and do so accurately, cooking experts are likely to rate cooking shows and do so accurately, but on average we may not get as accurate ratings on pairs with a sports and cooking video.

Thus one may wish to *separately* constrain *cross-group pairwise equal opportunity*:

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

and *within-group pairwise equal accuracy*:

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

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, but avoid trying to constrain cross-group ranking accuracy because we may not have confidence in cross-group ratings.

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 each group averaged over the other groups:

$$A_{G_i>:} = \kappa \text{ for some } \kappa \in [0, 1], \text{ for } i = 1, \ldots, K. \quad (7)$$

Statistical Parity

Our pairwise setup can also be used to define unsupervised fairness metrics. For any $i \neq j$, we define *pairwise statistical parity* as:

$$P\left(f(x) > f(x') \mid (x, y) \in G_i, (x', y') \in G_j\right) = \kappa. \quad (8)$$

A pairwise statistical parity 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* to be legitimate, or if raters are not expert enough to make useful cross-group comparisons.

Regression Pairwise Fairness Metrics

Consider the standard regression setting in which $f : \mathcal{X} \to \mathcal{Y}$ attempts to predict a regression label for each example. For most of the following proposed regression fairness metrics, we treat higher scores as more (or less) desirable, and we seek to control how often each group gets higher 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.

We again propose defining metrics on pairs of examples. This is not a ranking problem, so there are no queries—instead, given a training set of N examples, we compute pairwise metrics over all N^2 pairs. One can sample a *random subset* of pairs if N^2 is too large.

Regression Equal Opportunity

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.

Regression Equal Accuracy

Promoting *pairwise equal accuracy* as in (6) for regression requires that, for every group, the model should be equally faithful to the pairwise ranking of any two within-group examples. This is especially useful if the regression labels of different groups originate from different communities, and have different labeling distributions. For example, suppose that 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 minimizing MSE alone might cause the model training to over-focus on the classical music score examples, since the classical errors are likely to be larger and hence affect the MSE more.

Regression Statistical Parity

For regression, the pairwise statistical parity condition described in (8) requires, "Given two randomly drawn examples from two different groups, they are equally likely to have the higher score." One sufficient condition to guarantee pairwise statistical parity is to require the distribution of outputs $f(X)$ for a random input X to be the same for each of the protected groups. This condition can be enforced approximately by histogram matching the output distributions for different protected groups [e.g. Agarwal, Dudik, and Wu 2019].

Regression Symmetric 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.

Continuous Protected Features

Suppose we have a continuous or ordered protected feature Z; e.g. 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 constructing the pairs based on the ordering of the protected feature, rather on protected group membership. Specifically, we change (1) to the following *continuous attribute pairwise accuracies*:

$$A_> := P(f(x) > f(x') \mid y > y', z > z'), \quad (9)$$

$$A_< := P(f(x) > f(x') \mid y > y', z < z'), \quad (10)$$

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

$$A_> = A_<. \quad (11)$$

This requires, for example, that the model be equally accurate when the taller or shorter candidate should be higher ranked.[1]

Training for Pairwise Fairness

We show how one can use the pairwise fairness definitions to specify a training objective, and how to optimize these objectives. We formulate the training problem for ranking and cross-group equal opportunity, but the formulation and algorithms can be applied to any of the pairwise metrics.

Proposed Formulations Let $A_{G_i>G_j}(f)$ be defined by (1) for a ranking model $f : \mathcal{X} \to \mathbb{R}$. Let $AUC(f)$ be the overall pairwise accuracy. Let $\mathcal{F}$ be the class of models we are interested in. We formulate training with fairness goals as a *constrained optimization* with an allowed slack ϵ:

$$\max_{f \in \mathcal{F}} AUC(f)$$
$$\text{s.t.} \quad A_{G_i>G_j}(f) - A_{G_k>G_l}(f) \leq \epsilon \ \ \forall i \neq j, k \neq l. \quad (12)$$

Or one can pose the *robust optimization* problem:

$$\max_{f \in \mathcal{F}, \xi} \xi$$
$$\text{s.t.} \quad \xi \leq AUC(f), \xi \leq A_{G_i>G_j}(f) \ \ \forall i \neq j. \quad (13)$$

For regression problems, we replace AUC with MSE.

Optimization Algorithms Both the constrained and robust optimization formulations can be written in terms of *rate constraints* (Goh et al. 2016) on score differences. For example, we can re-write each pairwise accuracy term as a positive prediction rate on a subset of pairs:

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

where $\mathbb{I}$ is the usual indicator function and $\mathcal{S}_{ij} = \{((x,y),(x',y')) \,|\, y > y', (x,y) \in G_i, (x',y') \in G_j\}$. This enables us to adopt algorithms for binary fairness constraints to solve the optimization problems in (12) and (13).

In fact, *all* of the objective and constraint functions that we have considered can be handled out-of-the-box by the proxy-Lagrangian framework of Cotter, Jiang, and Sridharan; Cotter et al. (2019; 2019). Like other constrained optimization

[1] Similar to the pairwise ranking metrics, $A_<$ is the 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 $z < z'$:

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

Therefore, (11) equates both the TPR and the TNR for both sets of pairs, and specifies both equalized odds and equal opportunity.

approaches (Agarwal et al. 2018; Kearns et al. 2018), 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 min-max game, where one player minimizes over the model parameters, and the other player maximizes over a weighting λ on the constraint functions. Cotter, Jiang, and Sridharan (2019) use a no-regret optimization strategy for minimization over the model parameters, and a swap-regret optimization strategy for maximization over λ, with the indicators $\mathbb{I}$ replaced with hinge-based surrogates for the first player *only*. They prove that, under certain assumptions, their optimizers converge to a stochastic model that satisfies the specified constraints in expectation. In the Appendix, we present more details about the optimization approach and re-state their theoretical result for our setting.[2]

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.

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 (Goh et al. 2016; Narasimhan 2018; Cotter, Jiang, and Sridharan 2019; Cotter et al. 2019). For example, the goal of *demographic parity* (Dwork et al. 2012) is to ensure that the classifier's positive prediction rate is the same across all protected groups. Similarly, the *equal opportunity* metric (Hardt, Price, and Srebro 2016) requires that true positive rates 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* (Cotter et al. 2019). 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 (Zafar et al. 2015; Goh et al. 2016; Zafar et al. 2017; Donini et al. 2018; Agarwal et al. 2018; Cotter, Jiang, and Sridharan 2019; Cotter et al. 2019). Here, we extend this work to train ranking models and regression models with pairwise fairness constraints.

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. Zehlike et al.; Celis, Straszak, and Vishnoi; Biega, Gummadi, and Weikum; Singh and Joachims; Zehlike and Castillo; Singh and Joachims 2017; 2018; 2018; 2018; 2018; 2019]. 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. Singh and Joachims; Celis, Straszak, and Vishnoi; Zehlike and Castillo 2018; 2018; 2018], 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 (Biega, Gummadi, and Weikum 2018; Singh and Joachims 2018; 2019). Of these, some provide post-processing algorithms for re-ranking a given ranking (Biega, Gummadi, and Weikum 2018; Celis, Straszak, and Vishnoi 2018; Singh and Joachims 2018; 2019), while others, like us, learn a ranking model from scratch (Zehlike and Castillo 2018; Singh and Joachims 2019).

Pairwise Fairness Beutel et al. (2019) propose ranking pairwise fairness definitions equivalent to those we give in (1), (2) and (3). Their work focuses on ranking and on categorical protected groups, whereas we generalize these ideas to capture a wider variety of different statistical fairness notions, and generalize to regression and continuous protected features.

The training methodology is also very different. Beutel et al. (2019) propose adding a fixed regularization term to the training objective that measures the *correlation* between the residual between a clicked and unclicked item and the group memberships of the items. In contrast, we enable explicitly specifying any desired pairwise fairness constraints, and then directly enforce the desired pairwise fairness criterion using constrained optimization. Their approach is parameter-free, but only because it does not give the user any way to control the trade-off between fairness vs. accuracy.

Second, Beutel et al. consider 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 group must be penalized in order to satisfy the fairness constraints. A straightforward extension of the fixed regularization approach of Beutel et al. to multiple groups would have no hyperparameters to specify how much to weight each group. One could introduce separate weighting hyperparameters to weight each group's penalty, but then they would need to be tuned manually. The approach we propose does this tuning *automatically* to achieve the desired fairness constraints.

Finally, there are major experimental differences to Beutel et al.: 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.

Another recent work by Kallus and Zhou (2019) also provide pairwise fairness metrics based on AUC for bipartite ranking problems. However, they only consider categorical groups, whereas we also handle regression problems and continuous protected attributes. Further, their methodology is a post-processing approach that fits a monotonic transformation to an existing ranking model to optimize the specified metrics. In contrast, we provide a more flexible approach that enables optimizing the entire model by including the desired metrics as constrains during training.

Pinned AUC Pinned AUC is a fairness metric introduced by Dixon et al. (2018). With two protected groups, pinned AUC works by resampling the data such that each of the two groups make up 50% of the data, 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. (2019) demonstrate that pinned AUC,

[2]The appendix can be found in the full version of the paper: https://arxiv.org/pdf/1906.05330.pdf

as well as their proposed weighted pinned AUC metric, 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 different pairwise accuracies $A_{G_i>G_j}$ in (1). In our experiments, we compare against (a version of) the sampling-based approach of Dixon et al. (2018).

Fair Regression Defining fairness metrics in a regression setting is a challenging task, and has been studied for many years in the context of standardized testing [e.g. Hunter and Schmidt 1976]. Komiyama et al. (2018) consider the unfairness of a regressor in terms of the correlation between the output and a protected attribute. Pérez-Suay et al. (2017) 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.

Berk et al. (2017) propose regularizing linear regression models for the notion of fairness corresponding to the principle that *similar individuals receive similar outcomes* (Dwork et al. 2012). 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.

Agarwal, Dudik, and Wu (2019) propose a *bounded group loss* definition which requires that the regression error be within an allowable limit for each 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.

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 (Kearns et al. 2018). Fine-grained buckets raise statistical significance challenges, and coarse-grained buckets may raise unfairness issues due to how the lines between bins are drawn, and the lack of distinctions made between element within each bin. Raff, Sylvester, and Mills (2018) considered continuous protected features in their tree-growing criterion that addresses fairness. Kearns et al. (2018) focused on statistical parity-type constraints for continuous protected features for classification. Komiyama et al. (2018) controlled the correlation of the model output with protected variables (which may be continuous). Mary, Calauzènes, and Karoui (2019) propose a fairness criterion for continuous attributes based on the Rényi maximum correlation coefficient. *Counterfactual fairness* (Kusner et al. 2017; Pearl, Glymour, and Jewell 2016) 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.

Experiments

We illustrate our proposals on five ranking problems and two regression problems. We implement the constrained and robust optimization methods using the open-source Tensorflow constrained optimization toolbox of (Cotter, Jiang, and Sridharan 2019; Cotter et al. 2019). The datasets used are split randomly into training, validation and test sets in the ratio 1/2:1/4:1/4, with the validation set used to tune the relevant hyperparameters. For datasets with queries, we evaluate all metrics for individual queries and report the average across queries. For stochastic models, we report expectations over random draws of the scoring function f from the stochastic model.[3]

Pairwise Fairness for Ranking

We detail the comparisons and ranking problems.

Comparisons We compare against: (1) an adaptation of the *debiasing* scheme of Dixon et al. (2018) that optimizes a weighted pairwise accuracy, with the weights chosen to balance the relative label proportions within each group; (2) the recent non-pairwise ranking fairness approach by Singh and Joachims (2018) that re-ranks the scores of an unconstrained ranking model to satisfy a disparate impact constraint; (3) the post-processing pairwise fairness method of Kallus and Zhou (2019) that fits a monotone transform to an unconstrained model; and (4) the fixed regularization pairwise approach of Beutel et al. (2019) that like us incorporates the fairness goal into the model training. See Appendix for more details.

Simulated Ranking Data For this toy ranking task with two features, there are 5,000 queries, and each query has 11 candidates. For each query, we uniformly 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 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 $\mu_{-1,0} = [-1, 1]$, $\mu_{-1,1} = [-2, -1]$, $\mu_{+1,0} = [1, 0]$, $\mu_{+1,1} = [-1.5, 0.75]$, $\Sigma_{-1,0} = \Sigma_{-1,1} = \Sigma_{+1,0} = \mathbf{I}_2$ and $\Sigma_{+1,1} = 0.5\,\mathbf{I}_2$.

We train linear ranking functions $f : \mathbb{R}^2 \to \mathbb{R}$ and impose a *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 $\min\{A_{0>1}, A_{1>0}, AUC\}$. We also train an unconstrained model that optimizes AUC. Table 1 gives the test ranking accuracy, and the test pairwise fairness violations, measured as $|A_{0>1} - A_{1>0}|$. Only constrained optimization achieves the fairness goal, with robust optimization coming a close second. Figure 1(a)–(b) shows the 2×2 pairwise accuracy matrices. Constrained optimization satisfies the fairness constraint by lowering $A_{0>1}$ and improving $A_{1>0}$.

We also generate a second dataset with 3 groups, where the first two groups follow the same distribution as groups 0 and 1 above, and the third group examples are drawn from a Gaussian distribution $\mathcal{N}(\mu_{y,2}, \Sigma_{y,2})$ where $\mu_{-1,2} = [-1, 1], \mu_{+1,2} = [1.5, 0.5]$, and $\Sigma_{-1,2} = \Sigma_{+1,2} = \mathbf{I}_2$. We

[3]Code available at: https://github.com/google-research/google-research/tree/master/pairwise_fairness

Positive \ Negative	G_0	G_1
G_0	0.941	0.980
G_1	0.705	0.894

(a) Sim./Unconstrained

Positive \ Negative	G_0	G_1
G_0	0.854	0.900
G_1	0.907	0.927

(b) Sim./Constrained

G_0	G_1	G_2
0.713	0.703	0.729

(c) Sim./Constrained

G_0	G_1	G_2
0.882	0.876	0.949

(d) Sim./Robust

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

(e) Wiki/Unconstrained

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

(f) Wiki/Debiased

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

(g) Wiki/Constrained

Figure 1: Test pairwise accuracy matrices for Simulated ranking with 2 groups (a)–(b), row-based matrix averages $A_{0>:}$, $A_{1>:}$ and $A_{2>:}$ for Simulated ranking with 3 groups (c)–(d), and pairwise accuracy matrices for Wiki Talk Pages ranking (e)–(g).

Table 1: Test AUC (higher is better) 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, $|A_{G_0>:} - A_{G_1>:}|$ for marginal constraints, and $|A_> - A_<|$ for continuous protected attributes. Italicized text indicates strictly best between (Beutel et al. 2019) and constrained optimization.

Data	Groups	Uncons.	Debiased	S & J	K & Z	B et al.	Constr.	Robust
Sim.	0/1	**0.92** (0.28)	**0.92** (0.28)	0.88 (0.14)	0.91 (0.12)	0.84 (0.04)	*0.86* ***(0.01)***	0.86 (0.02)
Busns.	C/NC	**0.70** (0.06)	**0.70** (0.06)	–	0.66 (**0.00**)	0.69 (0.05)	0.68 (**0.00**)	0.68 (0.07)
Wiki	Term 'Gay'	**0.97** (0.10)	**0.97** (**0.01**)	–	**0.97** (0.04)	0.95 (**0.01**)	*0.96* ***(0.01)***	0.94 (0.02)
W3C	Gender	0.53 (0.96)	0.54 (0.90)	0.37 (0.85)	0.45 (0.65)	***0.55*** ***(0.09)***	0.54 (0.10)	0.54 (0.14)
Crime	Race %	**0.93** (0.18)	–	–	–	0.91 (0.10)	0.81 (**0.04**)	0.86 (**0.04**)

use the same number of queries and candidates as above, and assign the protected attribute z to $0, 1$, or 2 with probabilities $0.45, 0.1$, and 0.45 respectively. We impose the *marginal equal opportunity* fairness goal on this dataset in two different ways: *(i)* constraining $\max_{i \neq j} |A_{i>:} - A_{j>:}| \leq 0.01$ with constrained optimization, and *(ii)* optimizing $\min\{AUC, A_{0>:}, A_{1>:}, A_{2>:}\}$ with robust optimization. We show each group's row-marginal test accuracies in Figure 1(c)–(d). While robust optimization maximizes the minimum of the three marginals, constrained optimization yields a lower difference between the marginals (and does so at the cost of lower accuracies for the three groups). This is consistent with the two optimization problem set-ups: you get what you ask for.

We provide further results and an additional experiment with an *in-group* equal opportunity criterion in the appendix.

Business Matching This is a proprietary dataset from a large internet services company of ranked pairs of relevant and irrelevant businesses for different queries, for a total of 17,069 pairs. How well a query matches a candidate is represented by 41 features. We consider two protected groups, *chain* (C) businesses and *not chain* (NC) businesses. We define a candidate as a member of the *chain* group 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 if its query is not seeking a chain business and the candidate is a non-chain business. A candidate does not belong to either group if it is chain and the query is non-chain-seeking, or vice-versa.

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 chains as it at matching non-chains. With robust optimization, we maximize $\min\{A_{chain>:}, A_{non-chain>:}, AUC\}$. All methods trained a two-layer neural network model with 10 hidden nodes. As seen in Table 1, compared to the unconstrained approach, constrained optimization yields very low fairness violation, while only being marginally worse on the test AUC. The post-processing approach of (Kallus and Zhou 2019) also achieves a similar fairness metric, but with a lower AUC. (Singh and Joachims 2018) failed to produce feasible solutions for this dataset, we believe because there were very few pairs per query.

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 (Dixon et al. 2018)). This is a dataset where debiased weighting has been effective in learning fair, unbiased classification models (Dixon et al. 2018). 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. We consider the protected attribute defined by whether the term 'gay' appears in the comment. This is one of the many identity terms that Dixon et al. (2018) consider in their work. Among comments that have the term 'gay', 55% are labeled toxic, whereas among comments that do not have the term 'gay', only 9% are labeled toxic. We learn a convolutional neural network model with the same architecture used in Dixon et al. (2018).

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

Dataset	Prot. Group	Unconstrained	Beutel et al.	Constrained
Law	*Gender*	**0.142** (0.30)	0.167 (0.06)	*0.143* (***0.02***)
Crime	*Race %*	**0.021** (0.33)	0.033 (**0.02**)	0.028 (0.03)

High \ Low	Male	Female
Male	0.652	0.647
Female	0.666	0.655

Table 2: Left: Regression test MSE (lower is better) and pairwise fairness violation (in parenthesis), with italicized values indicating strictly best between last two columns. Right: Test pairwise accuracy matrix for *constrained* optimization on Law School dataset.

optimization. The results are shown in Table 1 and Figure 1(e)-(g). Among the cross-group errors, the unconstrained model is more likely to incorrectly rank a non-toxic comment with the term 'gay' over a toxic comment without the term. By balancing the label proportions, debiased weighting reduces the fairness violation considerably. Constrained optimization yields even lower fairness violation (0.010 vs. 0.014), but at the cost of a slightly lower test AUC. (Singh and Joachims 2018) could not be applied to this dataset as it did not have the required query-candidate structure.

W3C Experts Search We also evaluate our methods on the W3C Experts dataset, previously used to study disparate exposure in ranking (Zehlike and Castillo 2018). 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 (2018) 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 1, 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. 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. On this task, (Beutel et al. 2019) achieves the lowest fairness violation and the highest AUC.

The method of (Singh and Joachims 2018) performs poorly because the LP that it solves per query turns out to be infeasible for most queries in this dataset. Thus, to run this baseline, we extended their approach to have a per-query slack in their disparate impact constraints. This required a large slack for some queries, hurting the overall performance.

Communities and Crime (Continuous Groups) We next handle a *continuous protected attribute* in a ranking problem. We use the *Communities and Crime* dataset from UCI (Dua and Graff 2017) 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 (Cotter, Jiang, and Sridharan 2019), 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, and the methods of (Singh and Joachims 2018) and (Kallus and Zhou 2019), as they do not apply to continuous protected attributes. Adopting the continuous attribute equal opportunity criterion, we impose the constraint $|A_< - A_>| \leq 0.01$. We extend (Beutel et al. 2019) to optimize this pairwise metric. Table 1 shows the constrained and robust optimization methods reduce the fairness violation by more than half, at the cost of a lower test AUC.

Pairwise Fairness for Regression

We next present experiments on two regression problems.

We extend the set-up of Beutel et al. (2019) to also handle our proposed regression pairwise metrics, and compare to that. We do not use robust optimization as the squared error is not necessarily comparable with the regression pairwise metrics. The results are shown in Table 2.

Law School: This dataset (Wightman 1998) 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: $|A_{Female>Male} - A_{Male>Female}| \leq 0.01$. The constrained optimization approach successfully massively reduces the fairness violation compared to the unconstrained MSE-optimizing model, at only a small increase in MSE. It also performs strictly better than Beutel et al. (2019).

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 constrained approach yields a huge reduction in fairness violation, though at the cost of an increase in MSE.

Conclusions

We showed that pairwise fairness metrics can be intuitively defined to handle supervised and unsupervised notions of fairness, for ranking and regression, and for discrete and

continuous protected attributes. We also showed how pairwise fairness metrics can be incorporated into training using state-of-the-art constrained optimization solvers.

Experimentally, the different methods compared often produced different trade-offs between AUC (or MSE) and fairness, making it hard to judge one as strictly better than others. However, we showed that the proposed constrained optimization approach is the most flexible and direct of the strategies considered, and is very effective for achieving low pairwise fairness violations. The closest competitor to our proposals is the approach of Beutel et al. (2019), but out of the 4 cases in which one of these two methods strictly performed better (indicated in blue), ours was the best in 3 of the 4. Lastly, Kallus and Zhou (2019), being a post-processing method, is more restricted, and did not perform as well as the proposed approach that directly trains a model from scratch.

The key way one specifies pairwise fairness metrics is by the selection of which pairs to consider. Here, we focused on within-group and cross-group candidates. One could also bring in side information or condition on other features. For example, in the ranking setting, we might have side information about the presentation order that candidates for each query were shown to users when labeled, and this position information could be used to either select or weight candidate pairs. In the regression setting, we could assign weights to example pairs based on their label differences.

References

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

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

Berk, R.; Heidari, H.; Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; and Roth, A. 2017. A convex framework for fair regression. *FAT/ML*.

Beutel, A.; Chen, J.; Doshi, T.; Qian, H.; Wei, L.; Wu, Y.; Heldt, L.; Zhao, Z.; Hong, L.; Chi, E. H.; and Goodrow, C. 2019. Fairness in recommendation through pairwise experiments. *KDD*.

Biega, A. J.; Gummadi, K. P.; and Weikum, G. 2018. Equity of attention: Amortizing individual fairness in rankings. In *ACM SIGIR*.

Borkan, D.; Dixon, L.; Sorensen, J.; Thain, N.; and Vasserman, L. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *WWW*.

Celis, L. E.; Straszak, D.; and Vishnoi, N. K. 2018. Ranking with fairness constraints. In *ICALP*.

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

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

Cotter, A.; Jiang, H.; and Sridharan, K. 2019. Two-player games for efficient non-convex constrained optimization. In *ALT*.

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

Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J.; and Pontil, M. 2018. Empirical risk minimization under fairness constraints. *NeurIPS*.

Dua, D., and Graff, C. 2017. UCI machine learning repository.

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

Goh, G.; Cotter, A.; Gupta, M. R.; and Friedlander, M. P. 2016. Satisfying real-world goals with dataset constraints. In *NIPS*.

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

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

Kallus, N., and Zhou, A. 2019. The fairness of risk scores beyond classification: Bipartite ranking and the xAUC metric. *NeurIPS*.

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

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

Kusner, M.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. *NIPS*.

Liu, T. 2011. *Learning to Rank for Information Retrieval*. Springer.

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

Narasimhan, H. 2018. Learning with complex loss functions and constraints. In *AISTATS*.

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

Pérez-Suay, A.; Laparra, V.; Mateo-García, G.; Muñoz-Marí, J.; Gómez-Chova, L.; and Camps-Valls, G. 2017. Fair kernel learning. In *ECML PKDD*.

Raff, E.; Sylvester, J.; and Mills, S. 2018. Fair forests: Regularized tree induction to minimize model bias. *AIES*.

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

Singh, A., and Joachims, T. 2019. Policy learning for fairness in ranking. *NeurIPS*.

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

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

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and Gummadi, K. P. 2017. Fairness constraints: Mechanisms for fair classification. In *AISTATS*.

Zehlike, M., and Castillo, C. 2018. Reducing disparate exposure in ranking: A learning to rank approach. *arXiv preprint* arXiv:1805.08716.

Zehlike, M.; Bonchi, F.; Castillo, C.; Hajian, S.; Megahed, M.; and Baeza-Yates, R. A. 2017. FA*IR: A fair top-k ranking algorithm. In *CIKM*.