Fairness for Robust Learning to Rank

Omid Memarrast[1(✉)], Ashkan Rezaei[1], Rizal Fathony[2], and Brian Ziebart[1]

[1] University of Illinois Chicago, Chicago, IL 60607, USA
{omemar2,arezae4,bziebart}@uic.edu
[2] Grab, Jakarta 12430, Indonesia
rizal.fathony@grab.com

Abstract. While conventional ranking systems focus solely on maximizing the utility of the ranked items to users, fairness-aware ranking systems additionally try to balance the exposure based on different protected attributes such as gender or race. To achieve this type of group fairness for ranking, we derive a new ranking system from the first principles of distributional robustness. We formulate a minimax game between a player choosing a distribution over rankings to maximize utility while satisfying fairness constraints against an adversary seeking to minimize utility while matching statistics of the training data. Rather than maximizing utility and fairness for the specific training data, this approach efficiently produces robust utility and fairness for a much broader family of distributions of rankings that include the training data. We show that our approach provides better utility for highly fair rankings than existing baseline methods.

Keywords: Learning-to-rank · Fairness · Robustness

1 Introduction

Rankings often have social implications beyond the immediate utility they provide, since higher rankings provide opportunities for individuals and groups associated with the ranked items. As a consequence, biases in ranking systems, whether intentional or not, raise ethical concerns about their long-term economic and societal harming effect. Rankings that solely maximize utility or relevance can perpetuate existing societal biases that exist in training data whilst remaining oblivious to the societal detriment they cause by amplifying such biases [21].

Conventional ranking algorithms typically produce rankings to best serve the interests of those conducting searches by ordering the items by the probability of relevance so that utility to the users will be maximized [26]. Biased outcomes drawn by these models negatively impact items in marginalized protected groups in critical decision making systems such as hiring or housing where items compete for exposure and being unfair towards one group can lead to winner-takes-all dynamics that reinforce existing disparities [27].

H. Kashima et al. (Eds.): PAKDD 2023, LNAI 13935, pp. 544–556, 2023.
https://doi.org/10.1007/978-3-031-33374-3_43

Protected group definitions vary between different applications, and can include characteristics such as race, gender, religion, etc. In group fairness, algorithms divide the population into groups based on the protected attribute and guarantee the same treatment for members across groups. In ranking, this treatment can be evaluated using statistical metrics defined for measuring fairness. In this paper, we focus primarily on exposure-based group fairness measures. As a notable example, *demographic parity* (DP) in ranking is satisfied if the average exposure for both groups is equal in the top k ranks. As a motivating example, in Fig. 1 we consider two rankings based on items' true relevance and group membership. As a result of ranking 1, the highest utility is achieved, and fairness is ignored. In contrast, ranking 2 satisfies the demographic parity fairness constraint while still preserving high utility.

Fair ranking approaches seeking to provide group fairness properties can be categorized into *post-processing* and *in-processing* methods. *Post-processing* techniques are used to re-rank a given high utility ranking to incorporate fairness constraints while seeking to retain high utility [2,27]. These methods assume that true relevance labels are available and require other fairness-unaware learning methods (e.g., regression) to predict the true labels as a pre-processing step. Recovering from unfair regression based rankings in the re-ranking step may not be feasible in some circumstances [30].

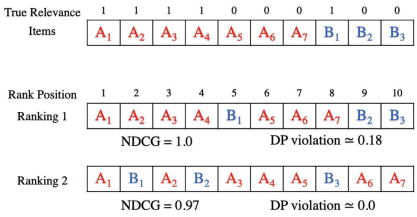

Fig. 1. Ranking 1 ignores fairness whereas Ranking 2 satisfies the demographic parity fairness constraint while only slightly decreasing the utility.

The fair ranking problem can also be addressed as an *in-processing*, learning-to-rank (LTR) task where the algorithm learns to maximize utility subject to fairness constraints from training data [28,31]. Our algorithm falls into this category. While providing a fairness-utility trade-off, fair LTR approaches need to be robust to outliers and noisy data. For example, the label of recidivism in the COMPAS dataset is regarded to be noisy [10]. This makes prediction while incorporating fairness constraints more difficult. With improved robustness properties, a fair LTR can achieve better utility for highly fair rankings, which results in a preferable utility-fairness trade-off.

In this paper, we derive a new LTR system based on the first principles of distributional robustness to provide both fairness and robustness to label noise. We formulate a minimax game with the *ranker player* choosing a distribution over rankings constrained to satisfy fairness requirements on the training samples while maximizing utility, and an *adversary player* choosing a distribution of item relevancies that minimizes utility while being similar to training data properties. Rather than narrowly optimizing the rankings for the specific training data, this approach produces rankings that provide utility and fairness robustly for a family of distributions that includes the training data.

We show that our approach is able to trade-off between utility and fairness much better at high levels of fairness than existing baseline methods. Furthermore, the robustness properties of our approach enable it to outperform existing baselines in the presence of varying degrees of label noise in the training data. To the best of our knowledge, this is the first distributionally robust fair LTR method.

2 Related Works

Fairness in Ranking. We can broadly group existing fair ranking approaches into various categories based on their notions of fairness. Metric-based works base their fairness constraints on statistical parity for pairwise ranking across item groups [1,15,20]. Several works argue that economic opportunities (e.g., exposure, clickthroughs, etc.) should be allocated on the basis of merit, not a winner-take-all strategy [2,9,27]. While our approach falls into this category, none of the existing techniques utilizes a distributionally robust approach to derive a fair LTR system like ours. As a result their performance degrades in the presence of training label noise, as we will show in our experiments.

There have also been recent studies that focus on other aspects of fair ranking. Several works have looked at fair ranking in the presence of noisy protected attributes [19]. Another line of research aims to select individuals distributed across different groups fairly when there is implicit group bias [6,16]. Recent studies have also investigated how uncertainty about protected attributes, labels, and other features of the machine learning model affect its fairness properties [12,22]. Contrary to this line of work, [29] takes into account the presence of uncertainty when estimating merits and defining a corresponding merit-based notion of fairness.

3 Preliminary

3.1 Probabilistic Ranking

To formulate the ranking task, we consider a dataset of ranking problems $\mathcal{D} = \{\mathcal{R}^i\}_{i=1}^N$ for N different queries, where each $\mathcal{R}^i = \{d_j\}_{j=1}^M$ is a candidate item set of size M for a single query. For every item d_j in this set, we denote $rel(d_j)$ as its corresponding relevance judgment. We denote the utility of a ranking (permutation) π for a single query as $\text{Util}(\pi)$. The optimization problem can be written as: $\pi^* = \text{argmax}_{\pi \in \Pi_{\text{fair}}} \text{Util}(\pi)$. Utility measures used for rankings are based on the relevance of the individual items being ranked for a particular ranking problem, $\mathcal{R}_{|\text{query}} = \{d_j\}_{j=1}^M$. For example, the Discounted Cumulative Gain (DCG) [14], which is a common evaluation measure for ranking systems that discounts the utility for lower-ranked items,

$$DCG(\pi) = \sum_{d_j \in \mathcal{R}} \frac{2^{rel(d_j)} - 1}{\log(1 + \pi_j)} \Rightarrow \text{Util}(\pi) = \sum_{j=1}^{M} u_j v_{\pi_j}, \tag{1}$$

is a member of the more general family of linear utility functions $u_j = 2^{rel(d_j)} - 1$ representing the utility of a single item d_j based on its relevance $rel(.)$ and $v_k = \frac{1}{\log(1+k)}$ providing the degree of attention that item d_j receives by being placed at rank k by permutation π, i.e., $\pi_{d_j} = k$.

The space of all permutations of items is exponential in the number of items, making naïve methods that find a utility-maximizing ranking subject to fairness constraints intractable. To overcome this problem, we consider a probabilistic ranking in which instead of a single ranking, a distribution over rankings is used. We define the probability of positioning item d_j at rank k as $P_{j,k}$. Then $\mathbf{P}$ constructs a doubly stochastic matrix of size $M \times M$ where entries in each row and each column must sum up to 1. By employing the idea of probabilistic ranking, we express the ranking utility in (1) as an expected utility of a probabilistic ranking:

$$\mathrm{U}(\mathbf{P}) = \sum_{j=1}^{M} \sum_{k=1}^{M} \mathbf{P}_{j,k}\, u_j\, v_k = \mathbf{u}^T \mathbf{P} \mathbf{v}, \tag{2}$$

which we equivalently express in a vectorized format where $\mathbf{u}$ and $\mathbf{v}$ are both column vectors of size M. Following [27], the fair ranking optimization can be expressed as a linear programming problem:

$$\max_{\mathbf{P} \in \Delta \cap \Gamma_{\text{fair}}} \mathbf{u}^T \mathbf{P} \mathbf{v} \quad \text{where:} \quad \Delta : \mathbf{P}\mathbf{1} = \mathbf{P}^\top \mathbf{1} = \mathbf{1},\ \ \mathbf{P}_{j,k} \geq 0,\ \ \forall_{1 \leq j,k \leq M} \tag{3}$$

and Γ_{fair} denotes any linear constraint set of the form $\mathbf{f}^\top \mathbf{P} \mathbf{g} = h$. Choosing $\mathbf{f}$ as the utility of items according to groups and $\mathbf{g}$ as the exposure of ranking position, enforces equality of exposure across protected groups. In contrast to [27], which uses this framework to re-rank the items to satisfy fairness constraints (i.e., a post-processing method), we extend this linear perspective to derive a *learning-to-rank* approach that learns to optimize utility and fairness simultaneously during training (i.e., an in-processing method).

Demographic parity of exposure, for a set of disjoint group members $G_1, \ldots, G_{|S|}$, requires that: $\frac{1}{|G_s|} \sum_{d_j \in G_s} \sum_{k=1}^{M} \mathbf{P}_{j,k} v_k = \frac{1}{|G_{s'}|} \sum_{d_j \in G_{s'}} \sum_{k=1}^{M} \mathbf{P}_{j,k} v_k, \forall s, s' \in S$.

In this paper, we assume binary groups and construct $\mathbf{f}_j = \frac{\mathbf{1}_{d_j \in G_s}}{|G_s|} - \frac{\mathbf{1}_{d_j \in G'_s}}{|G'_s|}$, which makes the constraint $\mathbf{f}^\top \mathbf{P} \mathbf{v} = 0$. For more than two groups, multiple pairwise constraints of this form can be enforced.

4 Methodology

We adopt a distributionally robust approach to the LTR problem by constructing a worst-case adversarial distribution on item utilities. We formulate the robust fair ranking construction as a minimax game between two players: a fair predictor $\mathbf{P}$ that makes a probabilistic prediction over the set of all possible rankings to maximize expected ranking utility; and an adversary $\mathbf{q}$ that approximates a probability distribution for the utility of items which minimizes the expected ranking utility. The adversary is additionally constrained to match the feature

moments of the empirical training distribution. Since we solve the problem for a given query, the query-dependent terms are omitted from the formulation for simplicity.

In our notation, we represent ranking items d by their feature representation $\mathbf{X} \in \mathbb{R}^{M \times L}$ as a matrix of M items with L features. For a given item set $\mathbf{X}$, the expected ranking utility of a probabilistic ranking $\mathbf{P}$ against a utility distribution $\mathbf{q}$ can be expressed as:

$$\mathrm{U}(\mathbf{X}, \mathbf{P}, \mathbf{q}) = \sum_{j=1}^{M} \mathbb{E}_{u_j|\mathbf{X}\sim\mathbf{q}} \left[u_j \mathbb{E}_{\pi_j|\mathbf{X}\sim\mathbf{P}} \left[v_{\pi_j} \right] \right]. \tag{4}$$

Then, the utility-maximizing optimization problem under fairness constraints can be formulated as:

Definition 1. *Given a training dataset of N ranking problems $\mathcal{D} = \{(\mathbf{X}^i, \mathbf{u}^i)\}_{i=1}^N$, with $\mathbf{u} \in \mathbb{R}^M$ being the true relevance and $\mathbf{X} \in \mathbb{R}^{M \times L}$ the feature representation of ranking problem of size M. The fair probabilistic ranking $\mathbf{P}(\pi) \in \mathbb{R}^{M \times M}$ in adversarial learning-to-rank learns a fair ranking that maximizes the worst-case ranking utility approximated by an adversary $\mathbf{q}(\check{\mathbf{u}})$, constrained to match the feature statistics of the training data.*

$$\max_{\mathbf{P}(\pi|\mathbf{X}) \in \Delta \cap \Gamma_{fair}} \min_{\mathbf{q}(\check{\mathbf{u}}|\mathbf{X})} \mathbb{E}_{\mathbf{X}\sim\widetilde{P}} \left[\mathrm{U}(\mathbf{X}, \mathbf{P}, \mathbf{q}) \right] \tag{5}$$

$$s.t. \quad \mathbb{E}_{\mathbf{X}\sim\widetilde{P}} \left[\sum_{j=1}^{M} \mathbb{E}_{\check{u}_j|\mathbf{X}\sim\mathbf{q}} \left[\check{u}_j \mathbf{X}_{j,:} \right] \right] = \mathbb{E}_{\mathbf{X},\mathbf{u}\sim\widetilde{P}} \left[\sum_{j=1}^{M} u_j \mathbf{X}_{j,:} \right] \tag{6}$$

where $\widetilde{P}$ denotes the empirical distribution over ranking dataset $\mathcal{D} = \{(\mathbf{X}^i, \mathbf{u}^i)\}_{i=1}^N$, $\check{\mathbf{u}}$ denotes the random variable for adversary relevance, and Δ denotes the set of doubly stochastic matrices.

This general adversarial formulation plays a foundational role in constructing probability models and prediction techniques [11,13]. This approach has been utilized to provide fair and robust predictions under covariate shift [25] as well as for constructing reliable predictors for fair log loss classification [24]. Similar to this line of work, our proposed approach imposes fairness constraints on predictor $\mathbf{P}$. Our formulation in Definition 1 accepts generic utility values. In our paper, we focus on binary utility, which is one of the common applications of the ranking problem, where the utility label indicates if a particular item is relevant or not. For the binary utility problem, the expected utility can be further simplified as:

$$\mathrm{U}(\mathbf{X}, \mathbf{P}, \mathbf{q}) = \sum_{j=1}^{M} \mathbb{E}_{u_j|\mathbf{X}\sim\mathbf{q}} \left[u_j \mathbb{E}_{\pi_j|\mathbf{X}\sim\mathbf{P}} \left[v_{\pi_j} \right] \right] = \sum_{j=1}^{M} \sum_{k=1}^{M} \mathbf{q}(u_j = 1|\mathbf{X}) \mathbf{P}(\pi_j = k|\mathbf{X}) v_k = \mathbf{q}^\top \mathbf{P} \mathbf{v},$$

where the entries in the vector $\mathbf{q}$ contains the relevance probability of item d_j. In the following sections, we use this vector notation to simplify the optimization formulation.

5 Optimization

We solve the constrained minimax formulation in Definition 1 in Lagrangian dual form, where we optimize the dual parameters $\theta \in \mathbb{R}^{L\times 1}$ for the feature matching constraint of L features by gradient descent. Rewriting the optimization in matrix notation yields:

$$\max_{\theta} \mathbb{E}_{\mathbf{x},\mathbf{u}\sim\widetilde{P}}\left[\max_{\mathbf{P}\in\Delta}\min_{0\leq\mathbf{q}\leq 1} \mathbf{q}^\top\mathbf{P}\mathbf{v} + \left\langle \mathbf{q}-\mathbf{u}, \sum\nolimits_l \theta_l \mathbf{X}_{:,l}\right\rangle\right] \text{ s.t. } \quad \mathbf{f}^\top\mathbf{P}\mathbf{v}=0, \quad (7)$$

where: $\mathbf{P}(\pi) \in \mathbb{R}^{M\times M}$ is a doubly stochastic matrix, and the value of cell $\mathbf{P}_{j,k}$ represents the probability that $\pi_j = k$; $\mathbf{u} \in \mathbb{R}^{M\times 1}$ is a vector of true labels whose j^{th} values is 1 when the item j is relevant to the query, i.e., $u_j = 1$ and 0 otherwise; $\mathbf{q} \in \mathbb{R}^{M\times 1}$ is a probability vector of the adversary's estimation of each item being relevant; $\mathbf{X}_{:,l} \in \mathbb{R}^{M\times 1}$ denotes the l^{th} feature of M samples; S is the set of protected attributes; and $\mathbf{v} \in \mathbb{R}^{M\times 1}$ is a vector containing the values of position bias function for each position. To denote the Frobenius inner product between two matrices $\langle ., . \rangle$ is used, i.e., $\langle A, B\rangle = \sum_{i,j} A_{i,j}B_{i,j}$.

For optimization purposes, using strong duality, we push the maximization over $\mathbf{q}$ to the outermost level in (7). Since the objective is non-smooth, for both $\mathbf{P}$ and $\mathbf{q}$, we add strongly convex prox-functions to make the objective smooth. Furthermore, to make our approach handle feature sampling error, we add a regularization penalty to the parameter θ. To apply (7) on training data, we replace empirical expectation with an average over all training samples. The new formulation is as follows:

$$\min_{\{0\leq\mathbf{q}^i\leq 1\}_{i=1}^N} \max_{\theta} \frac{1}{N}\sum_{i=1}^{N} \max_{\mathbf{P}^i\in\Delta}\left[{\mathbf{q}^i}^\top\mathbf{P}^i\mathbf{v}^i - \left\langle \mathbf{q}^i-\mathbf{u}^i, \sum\nolimits_l \theta_l \mathbf{X}^i_{:,l}\right\rangle\right.$$
$$\left. + \lambda{\mathbf{f}^i}^\top\mathbf{P}^i\mathbf{v}^i - \frac{\mu}{2}\left\|\mathbf{P}^i\right\|_F^2 + \frac{\mu}{2}\left\|\mathbf{q}^i\right\|_2^2\right] - \frac{\gamma}{2}\left\|\theta\right\|_2^2, \quad (8)$$

where superscript i is the i^{th} sample from N ranking problems in the training set. We denote λ, γ and μ as the fairness penalty parameter (which can be adjusted to obtain different trade-offs between fairness and utility, rather than strictly optimized), a regularization penalty parameter and a smoothing penalty parameter, respectively. The inner minimization over $\mathbf{P}$ and θ can be solved separately, given a fixed $\mathbf{q}$. The minimization over θ has a closed-form solution where the l^{th} element of θ^* is:

$$\theta_l^* = -\frac{1}{\gamma N}\sum_{i=1}^{N}\left\langle \mathbf{q}^i-\mathbf{u}^i, \mathbf{X}^i_{:,l}\right\rangle. \quad (9)$$

Independently from θ, we can solve the inner minimization over $\mathbf{P}$ for every training sample using a projection technique. The optimal $\mathbf{P}$ for i^{th} training sample (i.e., $\mathbf{P}^{i^*}$) is:

$$\mathbf{P}^{i^*} = \underset{\mathbf{P}^i \in \Delta}{\operatorname{argmax}} \ \mathbf{q}^{i^\top} \mathbf{P}^i \mathbf{v}^i + \lambda \mathbf{f}^{i^\top} \mathbf{P}^i \mathbf{v}^i - \frac{\mu}{2} \left\| \mathbf{P}^i \right\|_F^2$$

$$\mathbf{P}^{i^*} = \underset{\mathbf{P}^i \in \Delta}{\operatorname{argmin}} \ \frac{\mu}{2} \left\| \mathbf{P}^i - \frac{1}{\mu} (\mathbf{q}^i + \lambda \mathbf{f}^i) \mathbf{v}^{i^\top} \right\|_F^2 - \frac{1}{2\mu} \left\| \mathbf{q}^i \mathbf{v}^{i^\top} \right\|_F^2 . \tag{10}$$

As derived in (10), the minimization takes the form of $\min_{\mathbf{P} \geq 0} \|\mathbf{P} - \mathbf{R}\|_F^2$, and we can interpret this minimization as projecting matrix $\frac{1}{\mu}(\mathbf{q}^i + \lambda \mathbf{f}) \mathbf{v}^{i^\top}$ into the set of doubly-stochastic matrices. The projection from an arbitrary matrix $\mathbf{R}$ to the set of doubly-stochastic matrices can be solved using the ADMM projection algorithm [3]. Since each entry in $\mathbf{q}$ represents a probability, the outer optimization over $\mathbf{q}$ is solved using the L-BFGS-B algorithm with a bounded constraint of the probability simplex [4]. The algorithm optimizes the quadratic approximation of the objective function (using limited memory Quasi-Newton) on the convex set with each iteration. In each update step, a projection to the probability simplex is needed. Based on the above optimization, the adversary's optimal relevance probability $\mathbf{q}^*$ can be obtained. Following (9) we compute the θ^* over the optimal $\mathbf{q}^*$. Algorithm 1 shows the steps for training.

5.1 Inference and Runtime Analysis

For prediction, we use θ and μ learned from training data while performing the optimization in (8). After removing the constant terms, we solve a similar optimization problem for test data. That is:

Algorithm 1: The Fair-Robust LTR

Input: Training dataset $\mathcal{D} = \{(\mathbf{X}^i, \mathbf{u}^i)\}_{i=1}^N$, fairness penalty parameter λ.
Output: $\theta^*, \mathbf{P}^*, \mathbf{q}^*$
$\mathbf{q} \leftarrow$ random initialization;
repeat
  update θ by (9) with $\mathbf{q}$.
  update $\mathbf{P}$ by (10) with $\mathbf{q}$.
  update $\mathbf{q}$ by (8) with $\{\mathbf{P}, \theta\}$.
until *convergence*;

$$\min_{\{0 \leq \mathbf{q}^i \leq 1\}_{i=1}^{N^{\text{test}}}} \frac{1}{N^{\text{test}}} \sum_{i=1}^{N^{\text{test}}} \max_{\mathbf{P}^i \in \Delta} \left[\mathbf{q}^{i^\top} \mathbf{P}^i \mathbf{v}^i - \left\langle \mathbf{q}^i, \sum\nolimits_l \theta_l^* \mathbf{X}_{:,l}^i \right\rangle + \lambda \mathbf{P}^i \mathbf{v}^i - \frac{\mu}{2} \left\| \mathbf{P}^i \right\|_F^2 + \frac{\mu}{2} \left\| \mathbf{q}^i \right\|_2^2 \right],$$

where superscript i pertains to the i^{th} ranking problem in the test set of size N^{test}. We follow the steps for solving the optimization in training. There is no gradient learning of θ as in training, and true relevance labels ($\mathbf{u}$) are not used in inference. After convergence, we use the resulting $\mathbf{P}^*$ from the optimization to predict the ranking of items in the test set. We employ the Hungarian algorithm [17] to solve the problem of matching items to positions.

Runtime Analysis. Solving optimization in (8) involves running a projected gradient descent algorithm. In each iteration, it requires the computation of the gradient and the projection to box constraints. The box constraint projection's runtime is linear in terms of the number of variables, hence costing $\mathcal{O}(NM)$. The gradient computation requires solving for θ^*, which costs $\mathcal{O}(NML)$ from the dot product computations; and solving for $\mathbf{P}^*$, which can be posed as a doubly-stochastic matrix projection. We employ an ADMM algorithm to perform the projection to doubly stochastic matrix, which has linear convergence due to the strong convexity of the objective [7]. Each step inside the ADMM consists of M projections to M-element simplex, hence costing $\mathcal{O}(M^2)$ computations in total.

6 Experiments

In order to compare our proposed framework with existing fair LTR solutions, we use simulated and real-world datasets to carry out in-depth empirical evaluations. The learning task is to determine the feature function in the training based on the items' ground truth utilities and fairness constraints. At testing time, this feature function coupled with a penalty for fairness violation is used to determine the ranking for the items in the test set with maximum utility while satisfying fairness constraints.

6.1 Fairness Benchmark Datasets

Setup. We follow steps discussed in [28] to adapt German, Adult and COMPAS datasets to a LTR task. These datasets are inherently biased, making them viable alternatives for evaluation when no real world datasets exist for a fair LTR task. First, we split each dataset randomly into a disjoint train and test set. Then from each train/test set we construct a corresponding LTR train/test set. For each query, we sample randomly with replacement a set of 10 candidates each, representative of both relevant and irrelevant items, where on average four individuals are relevant. Each individual in the candidate set is a member of a group G_s based on its protected attribute. The training data consists of 500 ranking problems. We evaluate our learned model on 100 separate ranking problems serving as the test set. We repeat this process 10 times and report the 95% confidence interval in the results. The regularization constant γ and smoothing penalty parameter μ in (8) are chosen by 3-fold cross validation. We describe datasets used in our experiments:

- UCI `Adult`, census income dataset [8]. The goal is to predict whether income is above \$50K/yr on the basis of census results.
- The `COMPAS` criminal recidivism risk assessment dataset [18] is designed to predict whether a defendant is likely to reoffend based on criminal history.

Table 1. Dataset characteristics.

Dataset	n	Features	Attribute
`Adult`	45,222	12	Gender
`COMPAS`	6,167	10	Race
`German`	1,000	20	Gender

– UCI `German` dataset [8]. Based on personal information and credit history, the goal is to classify good and bad credit.

Table 1 shows the statistics of each dataset with their protected attributes.

Baseline Methods. To evaluate the performance of our model, we compare it against three different baselines that have similarities to and differences from our model: FAIR-PGRank [28] and DELTR [31] are in-processing, LTR methods, like ours; the Post-Processing method of [27] employs the fairness constraint formulation that we build our optimization framework based on. We also add a Random baseline that ranks items in each query randomly to give context to NDCG. We discuss baseline methods in more details[1]:

– **Post Processing** (Post-Proc) [27] To make a fair comparison with LTR approaches, we first learn a linear regression model using all query-item sets in the training data and predict the relevance of an item to a query in test set. Then, these estimated relevances are used as input to the linear program optimization described in [27] with a demographic parity constraint.
– **Fair Policy Ranking** (Fair-PGRank) [28] An end-to-end, in-processing LTR approach that uses a policy gradient method, directly optimizing for both utility and fairness measures.
– **Reducing Disparate Exposure** (DELTR) [31] An in-processing LTR method optimizing a weighted sum of a loss function and a fairness criterion. The loss function is a cross entropy designed for ranking [5] and fairness objective is a squared hinge loss based on disparate exposure.

Evaluation Metrics. We use the *normalized discounted cumulative gain* (NDCG) [14], as the utility measure. This is defined as: $NDCG(\pi) = DCG(\pi)/Z$, where Z is the DCG for ideal ranking and is used to normalize the ranking so that a perfect ranking would give a NDCG score of 1.

For the fairness evaluation in our approach we use *demographic parity* as our fairness violation metric which is based on disparity of average exposure across two groups:

$$\hat{D}_{group}(\mathbf{P}) = |\mathrm{Ex}(G_0|\mathbf{P}) - \mathrm{Ex}(G_1|\mathbf{P})|. \tag{11}$$

Results. Figure 2 shows the performance of our model (Fair-Robust) against baselines on the three benchmark datasets. We observe a trade-off between fairness and utility in both Fair-PGRank and Fair-Robust, i.e., as we increase the fairness penalty parameter (λ), demographic parity difference (as a measure of fairness violation) and NDCG both drop. While DELTR and Post-Proc achieve comparable NDCG when $\lambda = 0$, they fail to satisfy demographic parity

[1] We use the implementation from https://github.com/ashudeep/Fair-PGRank for all baselines.

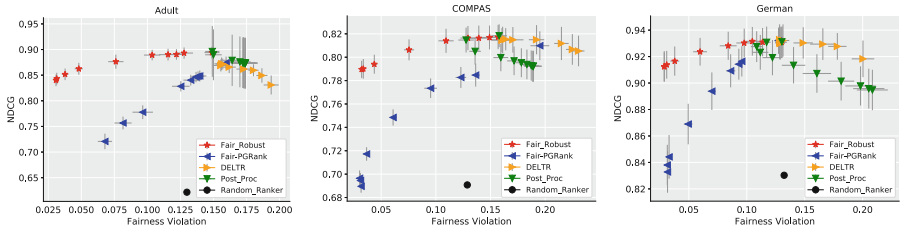

Fig. 2. Average *NDCG* versus average *difference of demographic parity* (DP) on test samples, for increasing degrees of fairness penalty λ in each method. FAIR-ROBUST: $\lambda \in [0, 20]$, FAIR-PGRANK: $\lambda \in [0, 20]$, DELTR: $\lambda \in [0, 10^6]$, POST-PROC: $\lambda \in [0, 0.2]$.

as we increase λ and are unable to provide a sufficient utility-fairness trade-off when high levels of fairness are desired.

In all three datasets, FAIR-ROBUST outperforms FAIR-PGRANK in terms of ranking utility when fairness is a priority. When comparing the utility-fairness trade-off between the two approaches, we observe that FAIR-ROBUST can retain higher NDCG in high levels of fairness and provides a preferable trade-off. One notable point is that, even in a noisy dataset like the `COMPAS` dataset, our approach performs better than other methods due to its robustness.

Robustness Test. One key benefit of our approach is its robustness to label noise in the learning process. This allows our method can be trained on data with noisy labels and outliers, and still perform well on the test data. To test this property, we repeat the previous experiment with noise added to the training data. After sampling rankings for the training and test sets, we randomly flip $x\%$ of the labels in each ranking problem in the training set. In our experiments, we test various amounts of noise in the training data where x can be 20%, 30%, or 40%. Figure 3 shows the results for robustness test. Similar to the previous experiment, we observe a trade-off between fairness and utility for FAIR-ROBUST. As the amount of the noise increases FAIR-PGRANK performs poorly and can't maintain its trade-off. Note that when $\lambda = 0$, FAIR-PGRANK still performs well but for other values of λ its NDCG gets close to random ranking.

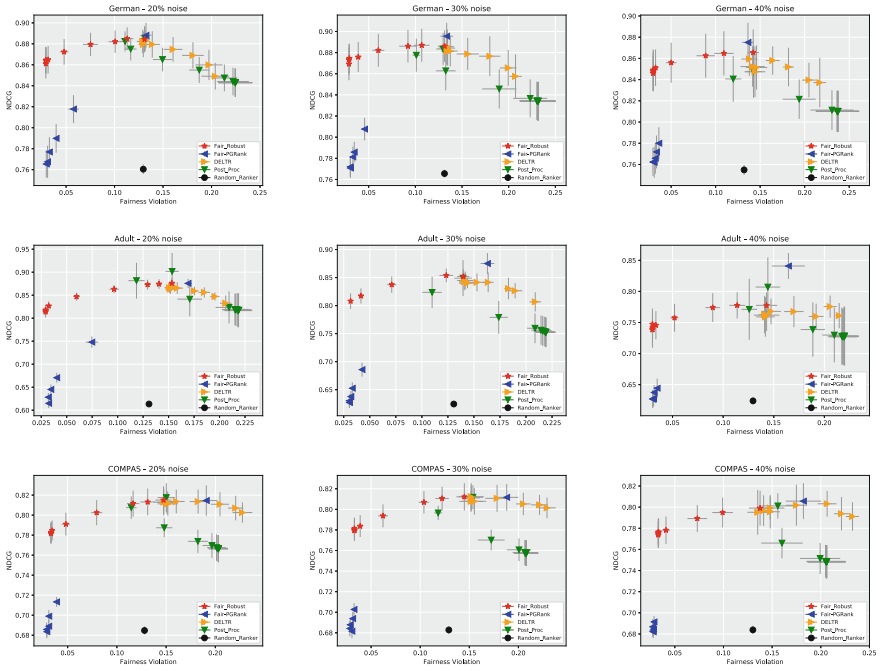

Fig. 3. Robustness test on `German`, `Adult` and `COMPAS` datasets with varying degrees of noise in the training data.

6.2 Microsoft Learning to Rank Dataset

Setup. In the previous experiments, we used datasets with inherent demographic biases but the LTR tasks were simulated and constructed from a classification task. In this experiment, we evaluate its performance on Microsoft's Learning to Rank dataset [23] which is a real world LTR dataset. We follow the steps discribed in [30] to pre-process the dataset. We compare our method to FAIR-PGRANK, as both methods are able to trade-off between fairness and utility. Additionally, we include a random baseline, which sorts each item in a query randomly, to give context to NDCG. Similar to the previous experiments, we use NDCG as the utility measure and *demographic parity* as our fairness violation metric, which is based on the disparity of average exposure across two groups.

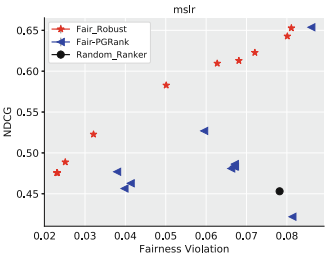

Fig. 4. *NDCG* versus *difference of demographic parity* for increasing degrees of fairness penalty λ in each method.

Results. Figure 4 shows the fairness and accuracy trade-off on the test set. With large fairness regularization, FAIR-PGRANK drops below a random rank-

ing in terms of NDCG, making it inconsistant. This plot shows that FAIR-ROBUST smoothly trades-off group fairness for NDCG. FAIR-PGRANK's NDCG and group exposure, on the other hand, deteriorate for increasing regularization strength, as [30] also observed.

7 Conclusions

In this paper, we developed a new LTR system that achieves fairness of exposure for protected groups while maximizing utility to the users. We show that our method is able to trade-off between utility and fairness much better at high levels of fairness than existing baseline methods. Our work addresses the problem of providing more robust fairness given a chosen fairness criterion, but does not answer the broader question of which fairness criterion is appropriate for a particular ranking application. More extensive evaluations based on incorporating other fairness metrics, such as disparate treatment, and generalization of this approach beyond binary utility are two important future directions.

Acknowledgements. This work was supported by the National Science Foundation Program on Fairness in AI in collaboration with Amazon under award No. 1939743.

References

1. Beutel, A., et al.: Fairness in recommendation ranking through pairwise comparisons. In: ACM SIGKDD International Conference on Knowledge Discovery & Data Mining(KDD), pp. 2212–2220 (2019)
2. Biega, A.J., Gummadi, K.P., Weikum, G.: Equity of attention: Amortizing individual fairness in rankings. In: The 41st International ACM Sigir Conference on Research and Development in Information Retrieval, pp. 405–414 (2018)
3. Boyd, S., Parikh, N., Chu, E.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Now Publishers Inc (2011)
4. Byrd, R.H., Lu, P., Nocedal, J., Zhu, C.: A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. **16**(5), 1190–1208 (1995)
5. Cao, Z., Qin, T., Liu, T.Y., Tsai, M.F., Li, H.: Learning to rank: from pairwise approach to listwise approach. In: Proceedings of the 24th International Conference on Machine Learning, pp. 129–136 (2007)
6. Celis, L.E., Mehrotra, A., Vishnoi, N.K.: Interventions for ranking in the presence of implicit bias. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 369–380 (2020)
7. Deng, W., Yin, W.: On the global and linear convergence of the generalized alternating direction method of multipliers. J. Sci. Comput. **66**(3) (2016)
8. Dheeru, D., Karra Taniskidou, E.: UCI machine learning repository (2017). http://archive.ics.uci.edu/ml
9. Diaz, F., Mitra, B., Ekstrand, M.D., Biega, A.J., Carterette, B.: Evaluating stochastic rankings with expected exposure. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 275–284 (2020)

10. Eckhouse, L.: Big data may be reinforcing racial bias in the criminal justice system. The Washington Post (2017)
11. Fathony, R., Liu, A., Asif, K., Ziebart, B.: Adversarial multiclass classification: A risk minimization perspective. In: NeurIPS (2016)
12. Ghosh, A., Dutt, R., Wilson, C.: When fair ranking meets uncertain inference. arXiv preprint arXiv:2105.02091 (2021)
13. Grünwald, P.D., Dawid, A.P.: Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. Ann. Stat. **32**, 1367–1433 (2004)
14. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Trans. Inform. Syst. (TOIS) **20**(4), 422–446 (2002)
15. Kallus, N., Zhou, A.: The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. In: Advances in Neural Information Processing Systems, pp. 3438–3448 (2019)
16. Kleinberg, J., Raghavan, M.: Selection problems in the presence of implicit bias. In: 9th Innovations in Theoretical Computer Science Conference (ITCS 2018) (2018)
17. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Res. Logist. Quart. **2**(1–2), 83–97 (1955)
18. Larson, J., Mattu, S., Kirchner, L., Angwin, J.: How we analyzed the compas recidivism algorithm. ProPublica 9 (2016)
19. Mehrotra, A., Celis, L.E.: Mitigating bias in set selection with noisy protected attributes. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 237–248 (2021)
20. Narasimhan, H., Cotter, A., Gupta, M., Wang, S.: Pairwise fairness for ranking and regression. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 5248–5255 (2020)
21. O'Neil, C.: Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books (2016)
22. Prost, F., et al.: Measuring model fairness under noisy covariates: A theoretical perspective. arXiv preprint arXiv:2105.09985 (2021)
23. Qin, T., Liu, T.Y.: Introducing letor 4.0 datasets. arXiv preprint arXiv:1306.2597 (2013)
24. Rezaei, A., Fathony, R., Memarrast, O., Ziebart, B.: Fairness for robust log loss classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 5511–5518 (2020)
25. Rezaei, A., Liu, A., Memarrast, O., Ziebart, B.D.: Robust fairness under covariate shift. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 9419–9427 (2021)
26. Robertson, S.E.: The probability ranking principle in ir. J. Document. **33**(4), 294–304 (1977)
27. Singh, A., Joachims, T.: Fairness of exposure in rankings. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2219–2228. ACM (2018)
28. Singh, A., Joachims, T.: Policy learning for fairness in ranking. Adv. Neural. Inf. Process. Syst. **32**, 5426–5436 (2019)
29. Singh, A., Kempe, D., Joachims, T.: Fairness in ranking under uncertainty. In: Advances in Neural Information Processing Systems, p. 34 (2021)
30. Yadav, H., Du, Z., Joachims, T.: Policy-gradient training of fair and unbiased ranking functions. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1044–1053 (2021)
31. Zehlike, M., Castillo, C.: Reducing disparate exposure in ranking: A learning to rank approach. In: Proceedings of The Web Conference 2020, pp. 2849–2855 (2020)