Fairness Through Regularization for Learning to Rank

Nikola Konstantinov [1] **Christoph H. Lampert** [1]

Abstract

Given the abundance of applications of ranking in recent years, addressing fairness concerns around automated ranking systems becomes necessary for increasing the trust among end-users. Previous work on fair ranking has mostly focused on application-specific fairness notions, often tailored to online advertising, and it rarely considers learning as part of the process.

In this work, we show how to transfer numerous fairness notions from binary classification to a learning to rank context. Our formalism allows us to design a method for incorporating fairness objectives with provable generalization guarantees. An extensive experimental evaluation shows that our method can improve ranking fairness substantially with no or only little loss of model quality.

1. Introduction

Ranking problems are abundant in many contemporary subfields of machine learning and artificial intelligence, including search, question answering, candidate/reviewer allocation, recommender systems and bid phrase suggestions (Manning et al., 2008). Decisions taken by such ranking systems affect our everyday life and this naturally leads to concerns about the fairness of ranking algorithms.

Indeed, ranking systems are typically designed to optimize for maximal utility and return the results most likely correct for each query (Robertson, 1977). This can have potentially harmful down-stream effects. For example, in 2015 Google became the target of heavy criticism after news reported that when searching for ”CEO” in Google’s image search, the first image of a women appeared only in the twelfth row, requiring two page scrolls to reach, and it actually did not show a real person but a Barbie doll (Lo, 2015). Similar problems exist in other ranking-based applications, such as product recommendations or online dating.

Potentially biased or otherwise undesirable results are particularly problematic in the *learning to rank (LTR)* setting (Liu, 2011; Mitra et al., 2018), where a machine learning model is trained to predict the relevances of the items for any query at test time. Training data for these systems is typically obtained from users interacting with another ranking system. Therefore, a biased selection of items can lead to disadvantageous winner-takes-it-all and rich-get-richer dynamics (Tsintzou et al., 2019; Tabibian et al., 2020)

A number of ranking-related fairness notions were proposed to make ranking systems more fair (Zehlike et al., 2017; Biega et al., 2018; Singh & Joachims, 2018). However, these were tailored to specific applications, such as online advertising. Some are also not well suited to the learning to rank situation, because their associated algorithms work by directly manipulating the order of returned items for a given query. In contrast, in the learning to rank setting one would hope that the system *learns to be fair*, such that a manipulation of the predicted relevance scores or the order of items are not necessary. This latter aspect brings up the problem of generalization for fairness: a ranking method could appear fair on the training data but be unfair at prediction time.

In this paper, we address these challenges and develop fairness-aware algorithms for learning to rank that provably generalize. To this end, we exploit connections between ranking and classification. Indeed, in contrast to fair learning in information retrieval, fair classification is a widely studied area where both the algorithmic and learning theoretic challenges of learning fair models are better understood (Barocas et al., 2019; Cotter et al., 2019; Woodworth et al., 2017). Importantly, many different notions of fairness have been proposed, which describe different properties that are desirable in various applications (Mehrabi et al., 2019).

We provide a formalism for translating such well-established and well-understood fairness notions from classification to ranking by phrasing the learning-to-rank problem as a binary classification problem for every separate query-item pair. We exemplify our approach on three fairness notions that emerge naturally in the ranking setting and correspond to popular concepts in the classification setting: *demographic parity*, *equalized odds*, and *equality of opportunity*. We then formulate corresponding *fairness regularization terms*, which can be incorporated with minor overhead into many standard learning to rank algorithms.

[1] IST Austria, Klosterneuburg, Austria. Correspondence to: Nikola Konstantinov <nikola.konstantinov@ist.ac.at>.

Besides its flexibility, another advantage of our approach is that it makes the task of fair learning to rank readily amendable to a learning-theoretic analysis. Specifically, we show generalization bounds for the three considered fairness notions, using a chromatic concentration bound for sums of dependent random variables (Janson, 2004) to overcome the challenge that different training samples for the same query are not statistically independent.

Finally, we demonstrate the practical usefulness of our approach for training fair models. Experiments on two ranking datasets confirm that training with a fairness regularizer can indeed yield models with greatly improved fairness at prediction time, and that the gains in fairness do not necessarily lead to a reduction of ranking quality.

2. Related work

Fairness in classification. Algorithmic fairness is well explored in the context of binary classification, see (Barocas et al., 2019) for a detailed introduction. In this work we show how to extend three popular group fairness notions – demographic parity, equalized odds and equality of opportunity (Hardt et al., 2016) – to the ranking setting. In principle, our formalism is applicable to other notions of group fairness, as well as individual (Dwork et al., 2012) and causal (Kusner et al., 2017) fairness notions. We defer studying these notions for ranking tasks to future work.

Fairness in ranking. Fairness in ranking has so far received less attention that fairness in classification. For an overview of recent techniques, see (Castillo, 2019). Most existing works concentrate on ranking-specific (single-purpose) fairness notions. One popular concept is *fairness of exposure* (Biega et al., 2018; Singh & Joachims, 2018; 2019; Yadav et al., 2019; Sapiezynski et al., 2019; Morik et al., 2020; Zehlike & Castillo, 2020; Gorantla et al., 2020). It states that exposure/attention received by a group of items or an individual item should be proportional to its utility. Other works aim at ensuring sufficient representation of items from different groups in the top-k positions of a ranking (Zehlike et al., 2017; Celis et al., 2018; Yang & Stoyanovich, 2017; Celis et al., 2020; Geyik et al., 2019). Besides group fairness also fair treatment of individuals has been studied in the context of ranking (Yang et al., 2019; Bower et al., 2021).

Among papers considering broader notions of fairness in ranking, Asudeh et al. (2019) design learning algorithms that can work with any fairness oracle. The framework however is limited to linear classifiers and the authors do not propose specific fairness notions. Singh & Joachims (2017) introduce a number of fair ranking definitions and draw parallels to equalized odds and demographic parity from fair classification. However, they do not provide a formal framework from studying the correspondence between the two setups, and they do not study how to optimize these measures in a learning to rank context. Moreover, their fairness measures concern fair rankings for a fixed query, which also holds for the causal fairness notion in (Wu et al., 2018). In contrast, our notion of ranking fairness is amortized across queries, similarly to (Biega et al., 2018).

Another related line of work is the one of pairwise fairness (Narasimhan et al., 2020; Beutel et al., 2019; Kuhlman et al., 2019). These works also describe ranking as a classification task in order to define fairness. However, the considered task is the proxy commonly employed by pairwise ranking methods, namely predicting which one of two items is more relevant that the other for a given query. In contrast, we define fairness in direct relation to the downstream task of deciding whether to return an item as relevant for a query or not. Kallus & Zhou (2019); Vogel et al. (2020) introduce fairness notions for bipartite ranking. These are also based on pairwise comparisons between points, but aim at learning fair continuous risks scores.

Overall, the main difference of our work to previous ones on ranking fairness is that we do not introduce a new fairness notion or algorithm. Instead, the formalism we introduce allows transferring existing fairness notions from classification to ranking. A second distinction is that only a minority of prior works considers fairness in the context of learning, and those who do usually propose new training techniques. Instead, the fairness regularizers we introduce can be combined with any existing training procedure that is formulatable as learning a score function by minimizing a cost function. Finally, no prior works provides generalization guarantees for fair ranking as we do.

Other related settings. For recommender systems, fairness can be studied with respect to the consumers/users (known as C-fairness) or with respect to the providers/items (known as P-fairness) (Burke, 2017). Steck (2018); Tsintzou et al. (2019) consider calibration and bias disparity within recommender systems with respect to recommended items. In (Burke et al., 2018; Farnadi et al., 2018; Zhu et al., 2018; Chakraborty et al., 2019; Peysakhovich & Kroer, 2019) various hybrid approaching for achieving both C-fairness and P-fairness are presented. However, these works are specific to collaborative filtering or tensor-based algorithms and do not carry over to approaches based on supervised learning.

Another related topic is the one of diversifying the output set of ranking system, see,e.g., (Radlinski et al., 2009). However, diversifying rankings generally has the goal of improving the user experience, not a fair treatment of items. A discussion on the relationship between fairness and ranking diversity can be found in (Singh & Joachims, 2018).

3. Preliminaries

In this section we introduce some background information on the learning to rank (LTR) task. A thorough introduction can be found, e.g., in (Liu, 2011; Mitra et al., 2018).

Learning to rank. Let $\mathcal{Q}$ be a set of possible *queries* to a ranking system, and let $\mathcal{D}$ be a set of *items* (historically *documents*) that are meant to be ranked according to their relevance for any query. Typically, we think of the query set as practically infinite, e.g. natural language phrases, whereas the item set is finite and fixed, for example, a database of products or customers. These are not fundamental constraints, though, and extensions are possible, e.g. items appearing or disappearing over time.

A dataset in the LTR setting typically has the form $S = \{(q_i, d^i_j, r^i_j)\}_{i\in[N],j\in[m_i]}$, i.e. for each of N queries, $q_1, \dots, q_N \in \mathcal{Q}$, a subset of the items $D_{q_i} = \{d^i_1, d^i_2, \dots, d^i_{m_i}\} \subset \mathcal{D}$ are annotated with binary labels $r^i_j = r(q_i, d^i_j) \in \{0, 1\}$ that indicate if item d^i_j is relevant to query q_i or not. In most real-world scenarios, m_i will be much smaller than $|\mathcal{D}|$, since it is typically impractical to determine the relevance of every item for a query.

The goal of learning to rank is to use a given training set to learn a *ranking procedure* that, for any future query, can return a set of items as well as their order. That is, the learner has to construct a *subset selection function*,

$$R : \mathcal{Q} \to \mathfrak{P}(\mathcal{D}), \tag{1}$$

where $\mathfrak{P}$ denotes the powerset operation, as well as an ordering of the predicted item set. In practice, both steps are typically combined by learning a *score function*, $s : \mathcal{Q}\times\mathcal{D} \to \mathbb{R}$. For any fixed q, $s(q, \cdot)$ induces a total ordering of $\mathcal{D}$, and the set of predicted items is obtained by thresholding or top-k prediction. The function s is usually learned by minimizing a loss function on the quality of the resulting ranking on the train data. Classic examples of this construction are SVM-Rank (Joachims, 2002) or WSABIE (Weston et al., 2011). Most other pointwise, pairwise and listwise methods can also be phrased in the above way, with differences mainly in the definition of the loss is defined and how the score function is learned numerically (Liu, 2011).

Evaluation measures. Many measures exist for evaluating the quality of a ranking system. Arguably the simplest is to measure the fraction of correctly predicted relevant items.

Definition 1. Let S be a test set in the format introduced above. For any query q_i, let $d^i_1, d^i_2, \dots$ be a ranking of the items in $\mathcal{D}_{q_i}$ with associated ground-truth values $r(q_i, d^i_j)$. Then, for any $k \in \mathbb{N} \setminus \{0\}$, the *precision at* k is defined as $\mathrm{P@k} = \frac{1}{N}\sum_{i=1}^N P@k(q_i)$ with

$$\mathrm{P@k}(q_i) = \frac{1}{k}\sum_{j=1}^{k} r(q_i, d^i_j). \tag{2}$$

For any k, the P@k value reflects only which items appear in the top-k list, but not their ordering. Furthermore, P@k is automatically small for datasets in which queries have only few relevant documents. To mitigate these shortcomings, one can add position-dependent weights and normalize by the score of a *best-possible* ranking.

Definition 2. In the same setting as for Definition 1, the *normalized discounted cumulative gain at* k is defined as $\mathrm{NDCG@}k = \frac{1}{N}\sum_{i=1}^N \mathrm{NDCG@}k(q_i)$ for

$$\mathrm{NDCG@}k(q_i) = \frac{\sum_{j=1}^{k} \frac{r(q_i,d^i_j)}{\log_2(j+1)}}{\sum_{j=1}^{l_i} \frac{1}{\log_2(j+1)}} \tag{3}$$

for $l_i = \min(k, K_i)$, where $K_i = |\{d \in \mathcal{D}_{q_i} : r(q_i, d) = 1\}|$ is the number of actually relevant items for query q_i. Queries with no relevant items are excluded from the average, as the measure is not well-defined for these.

4. Fairness in Learning-to-Rank

We now introduce our framework for group fairness in ranking. The main step is to exploit a correspondence between ranking and multi-label learning, a view that has previously been employed for practical tasks, e.g., in *extreme classification* (Bengio et al., 2019), but not –to our knowledge– to make LTR benefit from prior work on classification fairness.

Specifically, we study how a set of relevant items for any query can be selected in a fair way. Analogously to the discussion in Section 3, this originally means learning a *subset selection function* $R : \mathcal{Q} \to \mathfrak{P}(\mathcal{D})$, where $R(q)$ is the predicted set of selected items for a query q. The objects for which we want to impose fairness, the items, occur as outputs of the learned functions. This makes it hard to leverage fairness notions from the classification setting, where fairness is defined with respect to the input arguments.

We advocate an orthogonal viewpoint: for any fixed query q, we treat the items not as elements of the predictor's output, but as the inputs to a query-dependent classifier: $f_q : \mathcal{D} \to \{0, 1\}$, where $f_q(d) = 1$, if item d is should be returned for query q, and $f_q(d) = 0$ otherwise. As the query is a priori unknown, this means one ultimately has to find an *item selection function*

$$f : \mathcal{Q} \times \mathcal{D} \to \{0, 1\}. \tag{4}$$

While, of course, both views are completely equivalent, the latter one allows us to readily integrate notions of classification fairness into the learning to rank paradigm. In this work, we focus on the inclusion of *group fairness*, and leave the derivation of *individual fairness* (Dwork et al., 2012) or *counterfactual fairness* (Kusner et al., 2017) to future work.

Note that even though the item selection function $f(q, d)$ and the relevance label $r(q, d)$ have the same signature, their roles are different. r specifies if an item is relevant for a query or not. f indicates if the item should be returned as a result. These concepts differ when other aspects besides relevance are meant to influence the ranking, such as an upper bound on how many items can be retrieved per query or fairness and diversity considerations.

4.1. Group Fairness in Learning-to-Rank

Notions of group fairness in classification are typically based on an underlying probabilistic framework that allows statements about (conditional) independence relations (Barocas et al., 2019). The same is true in the ranking situation, where we assume $\mathbb{P} \in \mathcal{P}(\mathcal{Q} \times \mathcal{D} \times \{0,1\})$ to be an unknown but fixed distribution over query/document/relevance triplets. In the rest of our work, all statements about probabilities of events, denoted by Pr, will be with respect to $\mathbb{P}(q, d, r(q,d))$. Note that $\mathbb{P}$ characterizes only the marginal distribution of observing individual data points. It does not further specify how sets of many points, e.g. a training dataset, would be sampled. In particular, as we will discuss later, datasets for ranking tasks are typically not sampled i.i.d. from $\mathbb{P}$, but exhibit strong statistical dependencies.

Analogously to the situation of classification, we assume that any item $d \in \mathcal{D}$ has a *protected attribute*, $A(d)$, which denotes the group membership for which fairness should be ensured, e.g. the geographic origin when the items are products. In this work, we assume binary-valued protected attributes, but this is only for simplicity of presentation, not a fundamental limitation of our framework.

A plausible notion of fairness in the context of ranking is: **For any relevant item the probability of being included in the ranker's output should be independent of its protected attribute**. This intuition is easy to formulate in our formalism, resulting a direct analog of the *equality of opportunity* principle from fair classification (Hardt et al., 2016).

Definition 3 (**Equality of opportunity for LTR**). An item selection function $f : \mathcal{Q} \times \mathcal{D} \to \{0,1\}$ fulfills the *equality of opportunity* condition, if

$$\begin{aligned} &\Pr(f(q,d) = 1 | A(d) = 0, r(q,d) = 1) \\ &\quad = \Pr(f(q,d) = 1 | A(d) = 1, r(q,d) = 1), \end{aligned} \tag{5}$$

where $A(d)$ denotes the protected attribute of a document d.

The above definition provides a formal criterion what it means for a ranking system to be fair. In practice, however, a ranker will rarely achieve perfect fairness. Therefore, we also introduce a quantitative version of Definition 3 in the form of *a fairness deviation measure* (Woodworth et al., 2017; Williamson & Menon, 2019) that reports a ranking procedure's *amount of unfairness* (or *lack of fairness*) by means of its *mean difference score* (Calders & Verwer, 2010).

Definition 4. The *equality of opportunity (EOp) violation* of any item selection function, $f : \mathcal{Q} \times \mathcal{D} \to \{0,1\}$, is

$$\begin{aligned} \Gamma^{\text{EOp}}(f) = \Big| &\Pr(f(q,d)\!=\!1 | A(d)\!=\!0, r(q,d)\!=\!1) \\ &- \Pr(f(q,d)\!=\!1 | A(d)\!=\!1, r(q,d)\!=\!1) \Big|. \end{aligned} \tag{6}$$

Clearly, an item selection function, f, is fair in the sense of Definition 3 if and only if it fulfills $\Gamma^{\text{EOp}}(f) = 0$.

Other fairness measures. As discussed extensively in the literature, different notions of fairness are appropriate under different circumstances. For example, to check the *equality of opportunity* condition one needs to know which items are relevant for a query, and this can be problematic, e.g., if the available data itself exhibits a bias in this respect.

A major advantage of our formalism compared to prior ways for integrating fairness in ranking is that it is not partial to a specific fairness measure. Besides *equality of opportunity*, many other notions of group fairness can be expressed as easily by simply translating the corresponding expressions from the classification situation.

For example, one can avoid the problem of a data bias by simply demanding: **The probability of any item to be selected should be independent of its protected attribute** (i.e. disregarding its relevance to the specific query). In our formalism, this condition is a direct analog of the *demographic parity* criterion (Calders et al., 2009).

Definition 5 (**Demographic Parity for LTR**). An item selection function $f : \mathcal{Q} \times \mathcal{D} \to \{0,1\}$ fulfills the *demographic parity* condition, if

$$\Pr(f(q,d)\!=\!1 | A(d)\!=\!0) \!=\! \Pr(f(q,d)\!=\!1 | A(d)\!=\!1). \tag{7}$$

As associated quantitative measure we define the *demographic parity (DP) violation* of f as

$$\begin{aligned} \Gamma^{\text{DP}}(f) = \Big| &\Pr(f(q,d) = 1 | A(d) = 0) \\ &- \Pr(f(q,d) = 1 | A(d) = 1) \Big|. \end{aligned} \tag{8}$$

Another meaningful notion of fairness in ranking is: **The probability of any item to be selected should be independent of its protected attribute, individually for all relevant and for all irrelevant items.** This condition yields the ranking analog of the *equality odds* criterion (Hardt et al., 2016).

Definition 6 (**Equalized Odds for LTR**). An item selection function $f : \mathcal{Q} \times \mathcal{D} \to \{0,1\}$ fulfills the *equalized odds* condition, if for all $r \in \{0,1\}$:

$$\begin{aligned} &\Pr(f(q,d) = 1 | A(d) = 0, r(q,d) = r) \\ &\quad = \Pr(f(q,d) = 1 | A(d) = 1, r(q,d) = r) \end{aligned} \tag{9}$$

The *equalized odds (EOd) violation* of f is

$$\Gamma^{\text{EOd}}(f) = \frac{1}{2} \sum_{r \in \{0,1\}} \Big| \Pr(f(q,d)\!=\!1 | A(d)\!=\!0, r(q,d)\!=\!r)$$

$$- \Pr(f(q,d)=1|A(d)=1, r(q,d)=r)\Big|. \quad (10)$$

4.2. Training fair rankers

The above definitions do not only allow measuring the fairness of a fixed ranking system, but any of them can also be used to enforce the fairness of a learning to rank system during the training phase. For this, we create an empirical variant of the fairness violation measure and add it as a regularizer during the training step, thereby encouraging fairness of the learned ranker on the training set (Kamishima et al., 2011; Agarwal et al., 2018). For this construction to make sense, we have to answer two questions: *Can we solve the resulting optimization efficiently?* and *Does the inclusion of a regularizer generalize, i.e. ensure fairness not only at training time, but also on future predictions?* In rest of this section, we will answer the first question. The second question we will address in Section 4.3.

To allow for gradient-based optimization, we parametrize the binary-valued item selection function in a differentiable way using a real-valued score function $s : \mathcal{Q} \times \mathcal{D} \to [0,1]$, similar as introduced in Section 3. Our inspiration, however, comes from the classification setting, such as logistic regression, and we assume that s is not arbitrary real-valued, but that it parameterizes the probability that d is selected for q, i.e. $s(q,d) = \Pr(f(q,d) = 1)$.

Empirical fairness measures. For a given training set, S, in the format discussed in Section 3, we obtain empirical estimates of the previously introduced fairness violation measures. For any $a \in \{0,1\}$, $r \in \{0,1\}$, denote by S_a the subset of data points $(q,d,r(q,d))$ in S with $A(d) = a$, and by $S_{a,r}$ the subset of data points in S with $A(d) = a$ and $r(q,d) = r$.

Definition 7 (**Empirical fairness violation measures**). For any function $s : \mathcal{Q} \times \mathcal{D} \to [0,1]$, its *empirical equality of opportunity violation on a dataset S* is

$$\Gamma^{\text{EOp}}(s;S) = \Big|\frac{1}{|S_{0,1}|}\sum_{(q,d)\in S_{0,1}} s(q,d) - \frac{1}{|S_{1,1}|}\sum_{(q,d)\in S_{1,1}} s(q,d)\Big|. \quad (11)$$

The *empirical demographic parity violation of s on S* is

$$\Gamma^{\text{DP}}(s;S) = \Big|\frac{1}{|S_0|}\sum_{(q,d)\in S_0} s(q,d) - \frac{1}{|S_1|}\sum_{(q,d)\in S_1} s(q,d)\Big|. \quad (12)$$

and the *empirical equalized odds violation of s on S* is

$$\Gamma^{\text{EOd}}(s;S) = \frac{1}{2}\sum_{r\in\{0,1\}} \Big|\frac{1}{|S_{0,r}|}\sum_{(q,d)\in S_{0,1}} s(q,d) - \frac{1}{|S_{1,r}|}\sum_{(q,d)\in S_{1,1}} s(q,d)\Big|. \quad (13)$$

These expressions can be derived readily as approximations of the conditional probabilities of the individual fairness measures by fractions of the corresponding examples in S, by assuming that the marginal probability of any data point in S is $\mathbb{P}$, and inserting the assumed relation $s(p,q) = \Pr(f(p,q) = 1)$. Note that Definition 7 applies also to binary-valued functions, so it can also be used to evaluate the fairness of a learned item selection function on a dataset.

Learning with fairness regularization. Let $L(s,S)$ be any loss function ordinarily used to train an LTR system. Instead of optimizing solely this fairness-agnostic loss, we propose to optimize a fairness-aware regularized objective:

$$L^{\text{fair}}(s;S) = L(s,S) + \alpha\Gamma(s,S) \quad (14)$$

for $\alpha \geq 0$, where $\Gamma(s;S)$ is any of the empirical measures of fairness violation. The larger the value of α, the more the resulting rankers will take also the fairness of its decisions into account rather than just their utility.

As is typical for regularization constants, the optimal value of α depends on the specific application and the available data. In the case that constraints on the desired fairness of the system are given, e.g. the often cited *four-fifth rule* (Biddle, 2006), then a suitable value of α can be determined by classic model selection, e.g. using a validation set. In general, however, we expect the desired trade-off between utility and fairness to be influenced also by subjective factors, and we leave α as a free parameter in our discussion. However, as our experiments in Section 5 show, and as it has been observed in the context of classification as well, the relation between fairness and ranking quality is not necessarily adversarial. By choosing a reasonably small value of α, it can be possible to reduce unfairness while losing no or only a little bit of ranking quality.

Note that it is also possible to include more than one notion of fairness or fairness for multiple protected attributes simply by adding multiple corresponding regularization terms. We leave this aspect to future work, though.

Optimization. The fairness regularization terms, $\alpha\Gamma(s,S)$, are absolute values between differences of weighted sums over the score functions. Consequently, their values and their gradients can be computed efficiently using standard numerical frameworks. In large-scale settings, where ordinary gradient descent optimization is infeasible due to memory and computational limitations, the regularized objective (14) can also be optimized by stochastic gradient steps over mini-batches, as long as the unregularized loss function $L(s,S)$, supports this as well. The resulting per-batch gradient updates are not unbiased estimators of the full gradient, though, so the characteristics of the fairness notion changes depending on the batch size. For example, if batches were always formed of a single query with all associated documents, fairness would be enforced individually for each query, while the original objective enforces it averaged across all queries. In our experiments, however, we did not observe any deleterious effect of stochastic training when using a moderate batch size of 100.

4.3. Generalization

The method we propose for learning fair rankers works by enforcing fairness on a training set through a regularization term. In this section we show that –given enough data– this procedure will also ensure fairness at prediction time. Specifically, we prove a generalization bound by means of a uniform concentration argument, showing that the fairness on future decisions is bounded by the sum of the fairness on the training set and a complexity term, where the latter decreases monotonically towards zero with the number of queries in the training set. Our results are similar to the ones in (Woodworth et al., 2017) for the classification setting. However, in the case of ranking data there is additional dependence between the samples, which complicates the analysis and influences the complexity term.

Data generation process. To study the generalization properties of our fairness measures at training time versus prediction time, we first have to formally define the statistical properties of the training data. In this work, we assume the following data generation process which is consistent with the previously introduced structure of LTR datasets, with the only simplifying assumption that the item sets for all training queries are of equal size m.

For a given data distribution $\mathbb{P}(q, d, r)$, a dataset $S = \{(q_i, d^i_j, r^i_j)\}_{i\in[N], j\in[m]}$, is sampled as follows: 1) queries, $q_1, \dots, q_N$, are sampled i.i.d. from the marginal distribution $\mathbb{P}(q)$; 2) for each query q_i independently a set of items, $D_{q_i} = \{d^i_1, \dots, d^i_m\}$, is sampled *in an arbitrary way* with the only restriction that the marginal distribution of each individual d^i_j should be $\mathbb{P}(d|q_i)$; 3) for each pair (q_i, d^i_j) independently, the relevance r^i_j is sampled from $\mathbb{P}(r|q_i, d^i_j)$.

Note that each data point of the resulting training set has marginal distribution $\mathbb{P}$. Nevertheless, a lot of flexibility remains about how the actual items per query are chosen. In particular, the item set can have dependencies, let them be weak, such as avoiding repetitions, or strong, such as diversity constraints. While this choice of generating process complicates the theoretical analysis, we believe that it is necessary, because we want to make sure that real-world ranking data is covered, which typically is far from i.i.d.

We now characterize the generalization properties of the fairness regularizers. Let $\mathcal{F} \subset \{f : Q \times D \to \{0, 1\}\}$ be a set of item selection functions that make independent deterministic decisions per item (e.g., by thresholding a learned score function). Then, the following result holds.

Theorem 1. *Let S be a dataset sampled according to the above procedure with $2Nm > v$ for $v = VCdim(\mathcal{F})$. Let $P = \min_{r,a}\big(\mathbb{P}(r(q, d) = r \wedge A(d) = a)\big)$ and let $Q = \min_a \big(\mathbb{P}(A(d) = a)\big)$. Then, for any $\delta > 0$, each of the following inequalities holds with probability at least $1 - \delta$ over the sampling of S, uniformly for all $f \in \mathcal{F}$:*

$$\Gamma^{\text{EOp}}(f) \le \Gamma^{\text{EOp}}(f; S) + 8\sqrt{2\frac{v\log(\frac{2eNm}{v}) + \log(\frac{48}{\delta})}{NP^2}} \tag{15}$$

$$\Gamma^{\text{EOd}}(f) \le \Gamma^{\text{EOd}}(f; S) + 8\sqrt{2\frac{v\log(\frac{2eNm}{v}) + \log(\frac{48}{\delta})}{NP^2}} \tag{16}$$

$$\Gamma^{\text{DP}}(f) \le \Gamma^{\text{DP}}(f; S) + 8\sqrt{2\frac{v\log(\frac{2eNm}{v}) + \log(\frac{24}{\delta})}{NQ^2}} \tag{17}$$

Proof sketch. The proof consists of two parts. First, for any fixed item selection function a bound is shown on the gap between the conditional probabilities contributing to fairness measure and their empirical estimations. For this, we build on the technique of (Woodworth et al., 2017) for showing concentration of fairness quantities, in combination with the large deviations bounds for sums of dependent random variables of (Janson, 2004). Next, the bounds are extended to hold uniformly over the full hypothesis space by evoking a variant of the classic symmetrization argument (e.g. (Vapnik, 2013)), while carefully accounting for the dependence between the samples. A complete proof can be found in the supplementary material.

Discussion. Theorem 1 bounds the fairness violation on future data by the fairness on the training set plus an explicit complexity term, uniformly over all item selection functions. Consequently, any item selection function with low fairness violation on the training set will have a similarly low fairness violation on new data, provided that enough data was used for training.

The complexity term decreases like $\sqrt{\log N/N}$ as a function of the number of queries, N, which is the expected behavior for a VC-based bound. The same scaling behavior does not hold with respect to the number of items per query, m. This is unfortunate, but unavoidable, given the weak assumptions we make on the data generation process: because we do not restrict how the per-query item sets are created, each of them could simply consist of many copies of a single item. In that case, even arbitrary large m would provide only as much information as $m = 1$. In the current form, m appears even logarithmically in the numerator of the complexity term. We believe this to be an artifact of our proof technique, and expect that a more refined analysis will allow us to remove this dependence in the future. Note that for real data, we do expect larger m to be have a beneficial effect on generalization. This is the reason that we prefer to present the bound as it is in the theorem, i.e. with the empirical fairness estimated from all available data, rather than any alternative formulation, e.g. subsampling the training set to $m = 1$, which would recover an i.i.d. setting. Finding an assumption on the generating process of real-world LTR data that does allow bounds that decrease with respect to m is an interesting topic for future work.

5. Experiments

We report on some illustrative experiments to validate the practicality and performance of our method for training fair LTR systems, including a large-scale setting. Our emphasis lies on studying the interaction between model quality and fairness, as well as the effectiveness of our proposed method for optimizing both of these notions on real data. For space reasons, we only provide a high-level description of the experimental setting here. Technical details, e.g. on feature extraction, can be found in the supplemental material.

5.1. Datasets and experimental setup

We perform experiments on two datasets: the TREC 2019 Fairness data and MSMARCO. As a measure of ranking quality we use $\mathrm{NDCG@}k$ for $k \in \{1, 2, 3, 4, 5\}$, but also report results for $\mathrm{P@k}$ in the supplemental material. To quantify fairness, we evaluate the three different empirical measures of fairness violation for the corresponding top-k predictors.

TREC Fairness data. We use the training data of TREC 2019 Fairness track dataset (Biega et al., 2019). It consists of 652 real-world queries taken from the Semantic Scholar search engine, together with a set of scientific papers for each query and binary labels for the relevance of every query-paper pair. The average number of labeled papers per query is 7.1, out of which 3.4 are relevant on average. Because of the rather small number of queries, we use five-fold cross-validation to evaluate our method and report averages and standard errors across the folds.

As an exemplary protected attribute we use a proxy of the authors' seniority. We split the set of documents into two parts based on whether the mean of their authors' $i10$-index proxies (as provided in the TREC data) exceeds a threshold t or not. For $t \in \{3, 4, 5\}$ we get different amounts of group imbalance, with the minority group consisting of approximately $46\%, 26\%$ and 9% of all papers, respectively.

MSMARCO. We use the passage ranking dataset v2.1 of MSMARCO (Nguyen et al., 2016). It consists of approximately one million natural language questions, which serve as queries, associated sets of potentially relevant passages from Internet sources, and binary relevance labels for all provided query-document pairs. On average, there are 8.8 passages per question, and the average number of relevant ones is 0.65. For training and evaluation we use the default train-development split and report average and standard deviation over 10 random seeds.

To create a protected attribute, we split the passages into two groups based on their top-level domains, thinking of it as a proxy of the answers' geographic origin. Specifically, we split by `".com vs other"` (denoted by *com*) and by `".com/.org/.gov/.edu/.net vs other"` (denoted by *ext*). Their minority groups are of size 32% and 5% of all passages, respectively.

5.2. Learning to rank models

We adopt a classical pointwise learning to rank approach with a generalized linear score function, $s(d, q) = \langle\theta, \phi(q, d)\rangle$, for a predefined joint feature function, $\phi : \mathcal{Q} \times \mathcal{D} \to \mathbb{R}^D$ (see supplemental material). As loss function of ranking quality, $L(s, S)$, we use the squared loss between the relevance labels and the predictions of s over all data.

To optimize for both ranking quality and fairness, we train with a weighted loss, as in equation (14). For TREC we train all models by 1500 steps of gradient descent with a learning rate of 0.003. In the MSMARCO experiments we train with 5 epochs of SGD with a batch size of 100 queries and 10 passages per query and a learning rate of 0.0001.

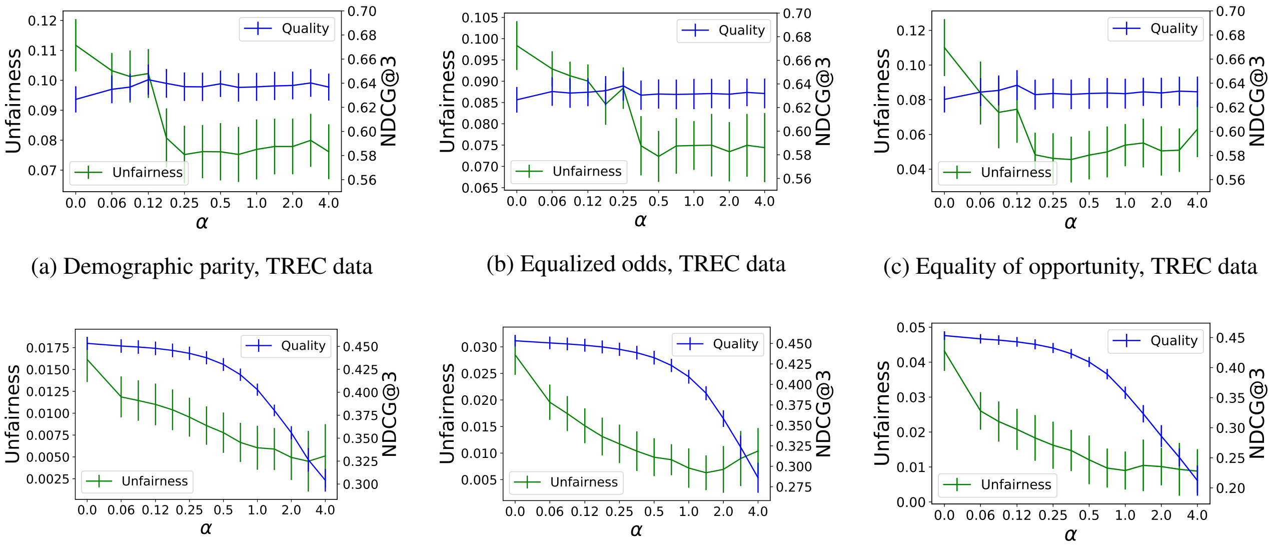
5.3. Results

Figure 1 show the results when imposing different amounts of the three discussed fairness notions in typical settings for TREC ($t = 3, k = 3$; top row) and MSMARCO (*com*, $k = 3$; bottom row). As one can see, our method is able to consistently improve fairness. For TREC, this comes at no loss in ranking quality (here NDCG). For MSMARCO the loss is quite small for small to medium values of α. As the figures shows, these observations are robust across the different amounts of regularization.

The possibility of increase the fairness of machine learning models without damaging their accuracy have been previously observed in the context of supervised learning (Wick et al., 2019).However, to the best of our knowledge, we are the first to observe this in a ranking context. Note that this effect is more expressed in the experiment on the TREC data than for MSMARCO. We hypothesize that this is because TREC on average has more relevant items per query. This means there is more flexibility which items to include in the top-k selection, thereby making room for increasing fairness while still retrieving sufficiently many relevant items.

We obtained very similar results also for the other experimental setups, such as different values of k, different protected attributes and a different measure of ranking quality. Plots for these can be found in the supplemental material.

Table 1 summarizes some of the results in a compact form. For different settings it reports the maximal reduction of



(a) Demographic parity, TREC data

(b) Equalized odds, TREC data

(c) Equality of opportunity, TREC data

(d) Demographic parity, MSMARCO

(e) Equalized odds, MSMARCO

(f) Equality of opportunity, MSMARCO

Figure 1: Results when learning fair rankers with our proposed regularization-based method: unfairness (left y-axes) and NDCG@3 ranking quality (right y-axes) after training with different regularization strengths $\alpha \in [0, 4]$.

the fairness violation measure over $\alpha \in [0, 4]$ for which the corresponding model's prediction quality is not significantly worse than for a model trained without a fairness regularizer (i.e. $\alpha = 0$). Here we say that a model is significantly worse than another if the difference of the mean quality values of the two models is larger than the sum of the standard errors/deviations around those averages (that is, if the error bars, as in Figure 1, would not intersect). It confirms that in all cases our proposed training method is able to strongly reduce the unfairness in the test time ranking without majorly damaging ranking quality.

Table 1: Maximal relative fairness increase without a significant decrease of ranking quality. See main text for details.

TREC		k 1	2	3	4	5	average over k
equality of opportunity	$t = 3$	35%	57%	59%	40%	32%	45%
	$t = 4$	54%	45%	46%	51%	35%	46%
	$t = 5$	65%	51%	47%	46%	27%	47%
demographic parity	$t = 3$	41%	32%	33%	14%	14%	27%
	$t = 4$	72%	40%	48%	38%	31%	46%
	$t = 5$	67%	40%	69%	51%	48%	55%
equalized odds	$t = 3$	20%	24%	26%	18%	16%	21%
	$t = 4$	32%	22%	32%	36%	23%	29%
	$t = 5$	28%	19%	32%	35%	22%	27%
average over settings		46%	37%	44%	36%	28%	38%

MSMARCO		k 1	2	3	4	5	average over k
equality of opportunity	*com*	73%	59%	58%	53%	51%	59%
	ext	29%	18%	20%	15%	27%	22%
demographic parity	*com*	36%	39%	41%	45%	51%	42%
	ext	10%	12%	17%	20%	26%	17%
equalized odds	*com*	75%	64%	64%	60%	61%	65%
	ext	28%	26%	30%	28%	40%	30%
average over settings		42%	36%	38%	37%	43%	39%

The results of two more experimental studies can be found in the supplemental material: first, an analog of Table 1 that gives not the maximal but the average improvement in fairness is reported. It confirms that the results are quite stable with respect to the choice of α, as was already visible in Figure 1. Second, we report results for the P@k measure. The results are almost identical to the ones for NDCG@k, confirming that the observed interaction between ranking quality and fairness is not tailored to a specific evaluation measure. Finally, we also report on results for a baseline model, where fairness is enforced for each query of the training set separately, similarly to (Singh & Joachims, 2017), rather than in average over all of them.

6. Conclusion

In this paper we introduced a framework for transferring supervised learning fairness notions to the context of learning to rank. The main step is to rephrase ranking as a collection of query-dependent classification problems. This viewpoint, while technically elementary, opens a wide range of possibilities for expanding the optimization methods and proof techniques from the fair classification literature to ranking and multi-label learning. In particular, we report the first—to the best of our knowledge—generalization bound for group fairness in the context of ranking.

Our experiments show that including a suitable regularizer during training can substantially improve the fairness of rankings with no or minor reduction in model quality. This effect seems even more pronounced than what had been observed in classification tasks, especially if the set of relevant items for any query is large. Therefore, we hypothesize that the multi-label nature of the ranking task naturally allows for more fairness without adverse effects on accuracy, and we deem making this intuition formal an interesting direction for future research.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *International Conference on Machine Learing (ICML)*, 2018.

Asudeh, A., Jagadish, H., Stoyanovich, J., and Das, G. Designing fair ranking schemes. In *International Conference on Management of Data (COMAD)*, 2019.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

Bengio, S., Dembczynski, K., Joachims, T., Kloft, M., and Varma, M. Extreme classification. In *Dagstuhl Reports 18291*. Schloss Dagstuhl – Leibniz Center for Informatics, 2019.

Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H., et al. Fairness in recommendation ranking through pairwise comparisons. In *Conference on Knowledge Discovery and Data Mining (KDD)*, 2019.

Biddle, D. *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Gower Publishing, 2006.

Biega, A. J., Gummadi, K. P., and Weikum, G. Equity of attention: Amortizing individual fairness in rankings. In *International Conference on Research and Development in Information Retrieval (SIGIR)*, 2018.

Biega, A. J., Diaz, F., Ekstrand, M. D., and Kohlmeier, S. Overview of the trec 2019 fair ranking track. In *The Twenty-Eighth Text REtrieval Conference (TREC 2019) Proceedings*, 2019.

Bonart, M. Fair ranking in academic search, 2019. URL `https://trec.nist.gov/pubs/trec28/papers/IR-Cologne.FR.pdf`.

Bower, A., Eftekhari, H., Yurochkin, M., and Sun, Y. Individually fair rankings. In *International Conference on Learning Representations (ICLR)*, 2021.

Burke, R. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093*, 2017.

Burke, R., Sonboli, N., and Ordonez-Gauger, A. Balanced neighborhoods for multi-sided fairness in recommendation. In *Conference on Fairness, Accountability and Transparency (FAccT)*, 2018.

Calders, T. and Verwer, S. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery (DMKD)*, 2010.

Calders, T., Kamiran, F., and Pechenizkiy, M. Building classifiers with independency constraints. In *International Conference on Data Mining Workshops (IDCMW)*, 2009.

Castillo, C. Fairness and transparency in ranking. In *International Conference on Research and Development in Information Retrieval (SIGIR)*, 2019.

Celis, L. E., Straszak, D., and Vishnoi, N. K. Ranking with fairness constraints. In *International Colloquium on Automata, Languages, and Programming (ICALP 2018)*. Schloss Dagstuhl – Leibniz Center for Informatics, 2018.

Celis, L. E., Mehrotra, A., and Vishnoi, N. K. Interventions for ranking in the presence of implicit bias. In *Conference on Fairness, Accountability and Transparency (FAccT)*, 2020.

Chakraborty, A., Patro, G. K., Ganguly, N., Gummadi, K. P., and Loiseau, P. Equality of voice: Towards fair representation in crowdsourced top-k recommendations. In *Conference on Fairness, Accountability and Transparency (FAccT)*, 2019.

Cotter, A., Gupta, M., Jiang, H., Srebro, N., Sridharan, K., Wang, S., Woodworth, B., and You, S. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In *International Conference on Machine Learing (ICML)*, 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference (ITCS)*, 2012.

Farnadi, G., Kouki, P., Thompson, S. K., Srinivasan, S., and Getoor, L. A fairness-aware hybrid recommender system. *arXiv preprint arXiv:1809.09030*, 2018.

Geyik, S. C., Ambler, S., and Kenthapadi, K. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Conference on Knowledge Discovery and Data Mining (KDD)*, 2019.

Gorantla, S., Deshpande, A., and Louis, A. Ranking for individual and group fairness simultaneously. *arXiv preprint arXiv:2010.06986*, 2020.

Han, S., Wang, X., Bendersky, M., and Najork, M. Learning-to-rank with bert in tf-ranking. *arXiv preprint arXiv:2004.08476*, 2020.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2016.

Janson, S. Large deviations for sums of partly dependent random variables. *Random Structures & Algorithms*, 24 (3):234–248, 2004.

Joachims, T. Optimizing search engines using clickthrough data. In *Conference on Knowledge Discovery and Data Mining (KDD)*, 2002.

Kallus, N. and Zhou, A. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Kamishima, T., Akaho, S., and Sakuma, J. Fairness-aware learning through regularization approach. In *International Conference on Data Mining Workshops (IDCMW)*, 2011.

Kuhlman, C., VanValkenburg, M., and Rundensteiner, E. Fare: Diagnostics for fair ranking using pairwise error metrics. In *International World Wide Web Conference (WWW)*, 2019.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

Liu, T.-Y. *Learning to rank for information retrieval*. Springer Science & Business Media, 2011.

Lo, D. When you Google image CEO, the first female photo on the results page is Barbie. `https://www.glamour.com/story/google-search-ceo`, 2015. Accessed: 2020-12-29.

Manning, C. D., Schütze, H., and Raghavan, P. *Introduction to information retrieval*. Cambridge University Press, 2008.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

Mitra, B., Craswell, N., et al. An introduction to neural information retrieval. *Foundations and Trends in Information Retrieval*, 2018.

Morik, M., Singh, A., Hong, J., and Joachims, T. Controlling fairness and bias in dynamic learning-to-rank. In *International Conference on Research and Development in Information Retrieval (SIGIR)*, 2020.

Narasimhan, H., Cotter, A., Gupta, M. R., and Wang, S. Pairwise fairness for ranking and regression. In *Conference on Artificial Intelligence (AAAI)*, 2020.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. Ms marco: A human-generated machine reading comprehension dataset. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2016.

Nogueira, R. and Cho, K. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*, 2019.

Peysakhovich, A. and Kroer, C. Fair division without disparate impact. *arXiv preprint arXiv:1906.02775*, 2019.

Radlinski, F., Bennett, P. N., Carterette, B., and Joachims, T. Redundancy, diversity and interdependent document relevance. In *International Conference on Research and Development in Information Retrieval (SIGIR)*, 2009.

Robertson, S. E. The probability ranking principle in ir. *Journal of documentation*, 33(4):294–304, 1977.

Sapiezynski, P., Zeng, W., E Robertson, R., Mislove, A., and Wilson, C. Quantifying the impact of user attentionon fair group representation in ranked lists. In *International World Wide Web Conference (WWW)*, 2019.

Singh, A. and Joachims, T. Equality of opportunity in rankings. In *Workshop on Prioritizing Online Content (WPOC) at NeurIPS*, 2017.

Singh, A. and Joachims, T. Fairness of exposure in rankings. In *Conference on Knowledge Discovery and Data Mining (KDD)*, 2018.

Singh, A. and Joachims, T. Policy learning for fairness in ranking. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Steck, H. Calibrated recommendations. In *Conference on Recommender Systems (RecSys)*, 2018.

Tabibian, B., Gómez, V., De, A., Schölkopf, B., and Rodriguez, M. G. On the design of consequential ranking algorithms. In *Uncertainty in Artificial Intelligence (UAI)*, 2020.

Tsintzou, V., Pitoura, E., and Tsaparas, P. Bias disparity in recommendation systems. In *Workshop on Recommendation in Multi-stakeholder Environments at RecSys*, 2019.

Vapnik, V. *The nature of statistical learning theory*. Springer, 2013.

Vogel, R., Bellet, A., and Clémençon, S. Learning fair scoring functions: Fairness definitions, algorithms and generalization bounds for bipartite ranking. *arXiv preprint arXiv:2002.08159*, 2020.

Weston, J., Bengio, S., and Usunier, N. Wsabie: Scaling up to large vocabulary image annotation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2011.

Wick, M., Panda, S., and Tristan, J.-B. Unlocking fairness: a trade-off revisited. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Williamson, R. and Menon, A. Fairness risk measures. In *International Conference on Machine Learing (ICML)*, 2019.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. Learning non-discriminatory predictors. In *Workshop on Computational Learning Theory (COLT)*, 2017.

Wu, Y., Zhang, L., and Wu, X. On discrimination discovery and removal in ranked data using causal graph. In *Conference on Knowledge Discovery and Data Mining (KDD)*, 2018.

Yadav, H., Du, Z., and Joachims, T. Fair learning-to-rank from implicit feedback. *arXiv preprint arXiv:1911.08054*, 2019.

Yang, K. and Stoyanovich, J. Measuring fairness in ranked outputs. In *Scientific and Statistical Database Management Conference (SSDBM)*, 2017.

Yang, K., Gkatzelis, V., and Stoyanovich, J. Balanced ranking with diversity constraints. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

Zehlike, M. and Castillo, C. Reducing disparate exposure in ranking: A learning to rank approach. In *International World Wide Web Conference (WWW)*, 2020.

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., and Baeza-Yates, R. Fa*ir: A fair top-k ranking algorithm. In *Conference on Information and Knowledge Management (CIKM)*, 2017.

Zhu, Z., Hu, X., and Caverlee, J. Fairness-aware tensor-based recommendation. In *Conference on Information and Knowledge Management (CIKM)*, 2018.

Supplemental Material

A. Proof of Theorem 1

In this section we present a complete proof of Theorem 1. To this end, we first introduce some classic definitions and concentration results for sums of dependent random variables from (Janson, 2004) in Section A.1. Next we show in Section A.2 how these can be used to derive large deviation bounds for the three fairness notions, given a fixed classifier. The proof is similar to the corresponding i.i.d. result of (Woodworth et al., 2017), however an application of the results from (Janson, 2004) is needed because of the dependence between the samples. Finally, in Section A.3 we show how these bounds can be made uniform over the hypothesis space by adapting the classic symmetrization argument (e.g. (Vapnik, 2013)) to a dependent data scenario.

A.1. Concentration inequalities for sums of dependent random variables

To deal with the dependence between the samples, we will use the following framework from (Janson, 2004). Let Y_α be a set of random variables, with α ranging over some index set $\mathcal{A}$. Let $X = \sum_{\alpha \in \mathcal{A}} Y_\alpha$. To derive concentration bounds for X, the following notions are useful:

Definition 8 ((Janson, 2004))**.** Given $\mathcal{A}$ and $\{Y_\alpha\}_{\alpha \in \mathcal{A}}$:

- A subset $\mathcal{A}' \subset \mathcal{A}$ is independent if the random variables $\{Y_\alpha\}_{\alpha \in \mathcal{A}'}$ are (jointly) independent.
- A family $\{\mathcal{A}_j\}_j$ is a cover of $\mathcal{A}$ if $\cup_j \mathcal{A}_j = \mathcal{A}$. A cover is proper if each set $\mathcal{A}_j$ is independent.
- $\chi(\mathcal{A})$ is the size of the smallest proper cover of $\mathcal{A}$, that is the smallest integer m, such that $\mathcal{A}$ can be written as the union of m independent subsets.

Then the following result holds, similar to the Hoeffding inequality, but accounting for the amount of dependence between the random variables $\{Y_\alpha\}_{\alpha \in \mathcal{A}}$:

Theorem 2 ((Janson, 2004))**.** *Let Y_α and X be as above, with $a_\alpha \leq Y_\alpha \leq b_\alpha$ for every $\alpha \in \mathcal{A}$, for some real numbers a_α and b_α. Then, for every $t > 0$:*

$$\mathbb{P}(X \geq \mathbb{E}(X) + t) \leq \exp\left(-2\frac{t^2}{\chi(\mathcal{A}) \sum_{\alpha \in \mathcal{A}} (b_\alpha - a_\alpha)^2}\right). \tag{18}$$

The same upper bound holds for $\mathbb{P}(X \leq \mathbb{E}(X) - t)$.

If instead one considers the mean of $\{Y_\alpha\}_{\alpha \in \mathcal{A}}$, namely $\bar{X} = \frac{1}{|\mathcal{A}|}\sum_{\alpha \in \mathcal{A}} Y_\alpha$, then the following holds:

$$\mathbb{P}(\bar{X} \geq \mathbb{E}(\bar{X}) + t) \leq \exp\left(-2\frac{t^2|\mathcal{A}|^2}{\chi(\mathcal{A}) \sum_{\alpha \in \mathcal{A}} (b_\alpha - a_\alpha)^2}\right). \tag{19}$$

Specifically, if the Y_α are Bernoulli random variables:

$$\mathbb{P}(\bar{X} \geq \mathbb{E}(\bar{X}) + t) \leq \exp\left(-2\frac{t^2|\mathcal{A}|}{\chi(\mathcal{A})}\right). \tag{20}$$

A.2. A non-uniform bound for equal odds

First we show a non-uniform Hoeffding-type bound for equal opportunity and equalized odds:

Lemma 1. *Fix $\delta \in (0,1)$ and a binary predictor $f : Q \times D \rightarrow \{0,1\}$. Suppose that $N > \frac{8\log(8/\delta)}{P^2}$, where $P = \min_{ar} \mathbb{P}(A(d) = a, r(q,d) = r)$, then:*

$$\mathbb{P}\left(|\Gamma^{EOp}(f,S) - \Gamma^{EOp}(f)| > 2\sqrt{\frac{\log(8/\delta)}{NP}}\right) \leq \delta. \tag{21}$$

and

$$\mathbb{P}\left(|\Gamma^{EOd}(f,S) - \Gamma^{EOd}(f)| > 2\sqrt{\frac{\log(16/\delta)}{NP}}\right) \leq \delta. \tag{22}$$

Proof. Denote by $I_{ar} = \{(i,j) : A(d^i_j) = a, r(q_i, d^i_j) = r\}$ the set of indexes of the training data for which the document belongs to the group a and the relevance of the query-document pair is r. Notice that I_{ar} is a random variable and that $|I_{ar}| = |S_{a,r}|$. We first bound the probability of a large deviation of

$$\gamma^S_{ar}(f) := \frac{1}{|I_{ar}|} \sum_{(i,j)\in I_{ar}} f(q_i, d^i_j)$$

from $\gamma_{ar}(f) := \mathbb{P}(f(q,d) = 1 | A(d) = a, r(q,d) = r)$, for each pair $r \in \{0,1\}, a \in \{0,1\}$. Since f is fixed here, we omit the dependence of $\gamma_{ar}(f), \gamma^S_{ar}(f), \Gamma^{\text{EOp}}(f), \Gamma^{\text{EOd}}(f)$, etc. on f for the rest of this proof.

For any fixed I_{ar}:

$$\mathbb{E}\left(\gamma^S_{ar} | I_{ar}\right) = \mathbb{E}\left(\frac{1}{|I_{ar}|} \sum_{(i,j)\in I_{ar}} f(q_i, d^i_j)\right) = \mathbb{P}(f(q,d) = 1 | A(d) = a, r(q,d) = r) = \gamma_{ar}(f), \tag{23}$$

since the marginal distribution of every $(q_i, d^i_j, r(q_i, d^i_j))$ is $\mathbb{P}$. It is also easy to see that if $\mathcal{A} = \{(i,j) : i \in [N], j \in [m]\}$ is the index set of the random variables $Y_{(i,j)} = f(q_i, d^i_j)$, then $\chi(\mathcal{A}) = m$. Therefore, for any fixed set $I_{ar} \subset \mathcal{A}$, we have $\chi(I_{ar}) \leq \chi(\mathcal{A}) = m$. Now conditional on I_{ar}:

$$\mathbb{E}(|\gamma^S_{ar} - \gamma_{ar}| > t | I_{ar}) = \mathbb{E}\left(\left|\frac{1}{|I_{ar}|} \sum_{(i,j)\in I_{ar}} f(q_i, d^i_j) - \gamma_{ar}\right| > t\right) \leq 2\exp\left(-2\frac{t^2 |I_{ar}|}{m}\right). \tag{24}$$

Similarly, $|I_{ar}| = \sum_{i\in[N]} \sum_{j\in[m]} \mathbb{1}(r(q_i, d^i_j) = r, A(d^i_j) = a)$ is the sum of Nm Bernoulli random variables indexed by $\mathcal{A} = \{(i,j)\}_{i\in[N], j\in[m]}$, such that $\chi(\mathcal{A}) = m$. Denote by $P_{ar} = \mathbb{P}(A(d) = a, r(q,d) = r)$ and recall the notation $P = \min_{ar} P_{ar}$. Then $\mathbb{E}(|I_{ar}|) = P_{ar}Nm$. Therefore,

$$\mathbb{P}\left(|I_{ar}| \leq P_{ar}Nm - t\right) \leq \exp\left(-2\frac{t^2}{Nm^2}\right).$$

Setting $t = P_{ar}Nm/2$, we obtain:

$$\mathbb{P}\left(|I_{ar}| \leq \frac{P_{ar}}{2}Nm\right) \leq \exp\left(-\frac{P^2_{ar}N}{2}\right). \tag{25}$$

Now assume that $N \geq \frac{2\log(8/\delta)}{P^2}$. Then for any $r \in \{0,1\}, a \in \{0,1\}$:

$$\begin{aligned}
\mathbb{P}(|\gamma^S_{ar} - \gamma_{ar}| > t) &= \sum_{I_{ar}} \mathbb{P}(|\gamma^S_{ar} - \gamma_{ar}| > t | I_{ar})\mathbb{P}(I_{ar}) \\
&\leq \mathbb{P}(|I_{ar}| \leq \frac{P_{ar}}{2}Nm) + \sum_{I_{ar}:|I_{ar}|\geq\frac{P_{ar}Nm}{2}} \mathbb{P}(|\gamma^S_{ar} - \gamma_{ar}| > t | I_{ar})\mathbb{P}(I_{ar}) \\
&\leq \exp\left(-\frac{P^2_{ar}N}{2}\right) + \sum_{I_{ar}:|I_{ar}|\geq\frac{P_{ar}Nm}{2}} 2\exp\left(-2\frac{t^2|I_{ar}|}{m}\right)\mathbb{P}(S_{ar}) \\
&\leq \frac{\delta}{8} + 2\exp\left(-t^2NP_{ar}\right).
\end{aligned}$$

The rest of the proof proceeds as in (Woodworth et al., 2017). For a fixed $r \in \{0,1\}$ the triangle law gives:

$$||\gamma^S_{0r} - \gamma^S_{1r}| - |\gamma_{0r} - \gamma_{1r}|| \leq |\gamma^S_{0r} - \gamma^S_{1r} - \gamma_{0r} + \gamma_{1r}| \leq |\gamma^S_{0r} - \gamma_{0r}| + |\gamma^S_{1r} - \gamma_{1r}|.$$

Therefore,

$$\begin{aligned}
\mathbb{P}(||\gamma^S_{0r} - \gamma^S_{1r}| - |\gamma_{0r} - \gamma_{1r}|| > 2t) &\leq \mathbb{P}(|\gamma^S_{0r} - \gamma_{0r}| + |\gamma^S_{1r} - \gamma_{1r}| > 2t) \\
&\leq \mathbb{P}((|\gamma^S_{0r} - \gamma_{0r}| > t) \vee (|\gamma^S_{1r} - \gamma_{1r}| > t)) \\
&\leq \mathbb{P}(|\gamma^S_{0r} - \gamma_{0r}| > t) + \mathbb{P}(|\gamma^S_{1r} - \gamma_{1r}| > t) \\
&\leq \frac{\delta}{4} + 4\exp(-t^2NP).
\end{aligned}$$

Setting $t = t_0 = \sqrt{\frac{\log(16/\delta)}{NP}}$ gives:

$$\mathbb{P}\left(||\gamma^S_{0r} - \gamma^S_{1r}| - |\gamma_{0r} - \gamma_{1r}|| > 2\sqrt{\frac{\log(16/\delta)}{NP}}\right) \leq \frac{\delta}{4} + 4\frac{\delta}{16} = \frac{\delta}{2}.$$

Setting $r = 1$ gives the first result.

For the second result, note that taking the union bound over $r \in \{0, 1\}$ shows that with probability at least $1 - \delta$ both $||\gamma_{00}^S - \gamma_{10}^S| - |\gamma_{00} - \gamma_{10}|| \leq 2t_0$ and $||\gamma_{01}^S - \gamma_{11}^S| - |\gamma_{01} - \gamma_{11}|| \leq 2t_0$ hold.

Under this event we have:

$$\begin{aligned}
|\Gamma^{\text{EOd}}(f, S) - \Gamma^{\text{EOd}}(f)| &= \left| \frac{1}{2} \left(|\gamma_{00}^S - \gamma_{10}^S| + |\gamma_{01}^S - \gamma_{11}^S| \right) - \frac{1}{2} \left(|\gamma_{00} - \gamma_{10}| + |\gamma_{01} - \gamma_{11}| \right) \right| \\
&= \left| \frac{1}{2} \left(|\gamma_{00}^S - \gamma_{10}^S| - |\gamma_{00} - \gamma_{10}| \right) + \frac{1}{2} \left(|\gamma_{01}^S - \gamma_{11}^S| - |\gamma_{01} - \gamma_{11}| \right) \right| \\
&\leq \frac{1}{2} \left| |\gamma_{00}^S - \gamma_{10}^S| - |\gamma_{00} - \gamma_{10}| \right| + \frac{1}{2} \left| |\gamma_{01}^S - \gamma_{11}^S| - |\gamma_{01} - \gamma_{11}| \right| \\
&\leq 2t_0
\end{aligned}$$

and hence the result follows. □

An identical argument, by conditioning on the values of the set $I_a = \{(i, j) : A(d_j^i) = a\}$ gives a similar result for demographic parity:

Lemma 2. *Fix $\delta \in (0, 1)$ and a binary predictor $f : Q \times D \to \{0, 1\}$. Suppose that $N > \frac{8 \log(8/\delta)}{Q^2}$, where $Q = \min_a \mathbb{P}(A(d) = a)$, then:*

$$\mathbb{P} \left(|\Gamma^{DP}(f, S) - \Gamma^{DP}(f)| > 2\sqrt{\frac{\log(8/\delta)}{NQ}} \right) \leq \delta. \tag{26}$$

A.3. A uniform bound for equal odds

Let $S' = \{(q_i', d_j'^i, r(q_i', d_j'^i))\}_{i \in [N], j \in [m]}$ be a ghost sample independent of S and also sampled via the same procedure as S, as described in the main body of the paper. In the proof of Lemma 1 we showed that for any classifier f and any $t \in (0, 1)$:

$$\mathbb{P} \left(|\Gamma^{\text{EOp}}(f) - \Gamma^{\text{EOp}}(f, S)| > 2t \right) \leq 2 \exp \left(-\frac{P^2 N}{2} \right) + 4 \exp \left(-\frac{t^2 N P}{2} \right) \leq 6 \exp \left(-\frac{t^2 N P^2}{2} \right) \tag{27}$$

$$\mathbb{P} \left(|\Gamma^{\text{EOd}}(f) - \Gamma^{\text{EOd}}(f, S)| > 2t \right) \leq 4 \exp \left(-\frac{P^2 N}{2} \right) + 8 \exp \left(-\frac{t^2 N P}{2} \right) \leq 12 \exp \left(-\frac{t^2 N P^2}{2} \right) \tag{28}$$

$$\tag{29}$$

Similarly, from the proof of Lemma 2

$$\mathbb{P} \left(|\Gamma^{\text{DP}}(f) - \Gamma^{\text{DP}}(f, S)| > 2t \right) \leq 2 \exp \left(-\frac{Q^2 N}{2} \right) + 4 \exp \left(-\frac{t^2 N Q}{2} \right) \leq 6 \exp \left(-\frac{t^2 N Q^2}{2} \right) \tag{30}$$

We will use these in particular to prove the following symmetrization lemma:

Lemma 3. *For any $1 > t \geq 4\sqrt{\frac{2 \log(12)}{NP^2}}$,*

$$\mathbb{P}_S \left(\sup_{f \in \mathcal{F}} (\Gamma^{EOp}(f) - \Gamma^{EOp}(f, S)) \geq t \right) \leq 2\mathbb{P}_{S,S'} \left(\sup_{f \in \mathcal{F}} (\Gamma^{EOp}(f, S') - \Gamma^{EOp}(f, S)) \geq t/2 \right). \tag{31}$$

For any $1 > t \geq 4\sqrt{\frac{2 \log(24)}{NP^2}}$:

$$\mathbb{P}_S \left(\sup_{f \in \mathcal{F}} (\Gamma^{EOd}(f) - \Gamma^{EOd}(f, S)) \geq t \right) \leq 2\mathbb{P}_{S,S'} \left(\sup_{f \in \mathcal{F}} (\Gamma^{EOd}(f, S') - \Gamma^{EOd}(f, S)) \geq t/2 \right). \tag{32}$$

For any $1 > t \geq 4\sqrt{\frac{2 \log(12)}{NQ^2}}$:

$$\mathbb{P}_S \left(\sup_{f \in \mathcal{F}} (\Gamma^{DP}(f) - \Gamma^{DP}(f, S)) \geq t \right) \leq 2\mathbb{P}_{S,S'} \left(\sup_{f \in \mathcal{F}} (\Gamma^{DP}(f, S') - \Gamma^{DP}(f, S)) \geq t/2 \right). \tag{33}$$

Proof. We show the result for the equal opportunity fairness measure, the rest follow in an identical manner.

Let f^* be the function achieving the supremum on the left-hand side [1]. Note that:

$$\begin{aligned}
&\mathbb{1}(\Gamma^{\mathrm{EOp}}(f^*) - \Gamma^{\mathrm{EOp}}(f^*, S) \geq t)\mathbb{1}(\Gamma^{\mathrm{EOp}}(f^*) - \Gamma^{\mathrm{EOp}}(f^*, S') < t/2) \\
&\qquad = \mathbb{1}(\Gamma^{\mathrm{EOp}}(f^*) - \Gamma^{\mathrm{EOp}}(f^*, S) \geq t \wedge \Gamma^{\mathrm{EOp}}(f^*, S') - \Gamma^{\mathrm{EOp}}(f^*) > -t/2) \\
&\qquad \leq \mathbb{1}(\Gamma^{\mathrm{EOp}}(f^*, S') - \Gamma^{\mathrm{EOp}}(f^*, S) > t/2).
\end{aligned}$$

Taking expectation with respect to S':

$$\mathbb{1}(\Gamma^{\mathrm{EOp}}(f^*) - \Gamma^{\mathrm{EOp}}(f^*, S) \geq t)\mathbb{P}_{S'}(\Gamma^{\mathrm{EOp}}(f^*) - \Gamma^{\mathrm{EOp}}(f^*, S') < t/2) \leq \mathbb{P}_{S'}(\Gamma^{\mathrm{EOp}}(f^*, S') - \Gamma^{\mathrm{EOp}}(f^*, S) > t/2).$$

Now using (27):

$$\mathbb{P}_{S'}(\Gamma^{\mathrm{EOp}}(f^*) - \Gamma^{\mathrm{EOp}}(f^*, S') \geq t/2) \leq 6 \exp\left(-\frac{t^2 N P^2}{32}\right) \leq \frac{1}{2},$$

so:

$$\frac{1}{2}\mathbb{1}(\Gamma^{\mathrm{EOp}}(f^*) - \Gamma^{\mathrm{EOp}}(f^*, S) \geq t) \leq \mathbb{P}_{S'}(\Gamma^{\mathrm{EOp}}(f^*, S') - \Gamma^{\mathrm{EOp}}(f^*, S) > t/2).$$

Taking expectation with respect to S:

$$\begin{aligned}
\mathbb{P}_S(\Gamma^{\mathrm{EOp}}(f^*) - \Gamma^{\mathrm{EOp}}(f^*, S) \geq t) &\leq 2\mathbb{P}_{S,S'}(\Gamma^{\mathrm{EOp}}(f^*, S') - \Gamma^{\mathrm{EOp}}(f^*, S) > t/2) \\
&\leq 2\mathbb{P}_{S,S'}(\sup_{f \in \mathcal{F}}(\Gamma^{\mathrm{EOp}}(f, S') - \Gamma^{\mathrm{EOp}}(f, S)) \geq t/2).
\end{aligned}$$

□

Given a set of n input datapoints $z_1, \ldots, z_n$ with $z_i = (q_i, d_i, r(q_i, d_i))$, consider:

$$\mathcal{F}_{z_1,\ldots,z_n} = \{(f(q_1, d_1), \ldots, f(q_n, d_n)) : f \in \mathcal{F}\} \tag{34}$$

Then the growth function of $\mathcal{F}$ is defined as:

$$S_{\mathcal{F}}(n) = \sup_{(z_1,\ldots,z_n)} |\mathcal{F}_{z_1,\ldots,z_n}| \tag{35}$$

We can now present a proof of Theorem 1:

Theorem 1. *Suppose that $v = VC(\mathcal{F}) \geq 1$ and that $2Nm > v$. Then for any $\delta \in (0, 1)$:*

$$\mathbb{P}_S\left(\sup_{f \in \mathcal{F}}(\Gamma^{EOp}(f) - \Gamma^{EOp}(f, S)) \geq 8\sqrt{2\frac{d\log(\frac{2eNm}{d}) + \log(\frac{24}{\delta})}{NP^2}}\right) \leq \delta \tag{36}$$

$$\mathbb{P}_S\left(\sup_{f \in \mathcal{F}}(\Gamma^{DP}(f) - \Gamma^{DP}(f, S)) \geq 8\sqrt{2\frac{d\log(\frac{2eNm}{d}) + \log(\frac{24}{\delta})}{NQ^2}}\right) \leq \delta \tag{37}$$

$$\mathbb{P}_S\left(\sup_{f \in \mathcal{F}}(\Gamma^{EOd}(f) - \Gamma^{EOd}(f, S)) \geq 8\sqrt{2\frac{d\log(\frac{2eNm}{d}) + \log(\frac{48}{\delta})}{NP^2}}\right) \leq \delta \tag{38}$$

Proof. Again we present the proof for equal opportunity, with the other inequalities following in an identical manner.

Note that given sets S and S', the values of $\Gamma^{\mathrm{EOp}}(f, S)$ and $\Gamma^{\mathrm{EOp}}(f, S')$ are completely determined by the values of f on S and S' respectively. Therefore, for any $t \in \left(4\sqrt{\frac{2\log(12)}{NP^2}}, 1\right)$ using Lemma 3 and the union bound:

$$\begin{aligned}
\mathbb{P}_S\left(\sup_{f \in \mathcal{F}}(\Gamma^{\mathrm{EOp}}(f) - \Gamma^{\mathrm{EOp}}(f, S)) \geq t\right) &\leq 2\mathbb{P}_{S,S'}\left(\sup_{f \in \mathcal{F}}(\Gamma^{\mathrm{EOp}}(f, S') - \Gamma^{\mathrm{EOp}}(f, S)) \geq t/2\right) \\
&\leq 2S_{\mathcal{F}}(2Nm)\mathbb{P}_{S,S'}\left(\Gamma^{\mathrm{EOp}}(f, S') - \Gamma^{\mathrm{EOp}}(f, S) \geq t/2\right) \\
&\leq 2S_{\mathcal{F}}(2Nm)\mathbb{P}_{S,S'}\left(|\Gamma^{\mathrm{EOp}}(f, S') - \Gamma^{\mathrm{EOp}}(f)| \geq t/4 \vee |\Gamma^{\mathrm{EOp}}(f) - \Gamma^{\mathrm{EOp}}(f, S)| \geq t/4\right) \\
&\leq 4S_{\mathcal{F}}(2Nm)\mathbb{P}_S\left(|\Gamma^{\mathrm{EOp}}(f) - \Gamma^{\mathrm{EOp}}(f, S)| \geq t/4\right)
\end{aligned}$$

[1] If the supremum is not attained, this argument can be repeated for each element of a sequence of classifiers approaching the supremum

$$\leq 24 S_{\mathcal{F}}(2Nm) \exp\left(-\frac{t^2 N P^2}{128}\right)$$

In particular, if $d = VC(\mathcal{F})$, by Sauer's lemma $S_{\mathcal{F}}(2Nm) \leq \left(\frac{2eNm}{d}\right)^d$ whenever $2Nm > d$, so:

$$\mathbb{P}_S\left(\sup_{f\in\mathcal{F}}(\Gamma^{\text{EOp}}(f) - \Gamma^{\text{EOp}}(f,S)) \geq t\right) \leq 24\left(\frac{2eNm}{d}\right)^d \exp\left(-\frac{t^2NP^2}{128}\right)$$

It follows that:

$$\mathbb{P}_S\left(\sup_{f\in\mathcal{F}}(\Gamma^{\text{EOp}}(f) - \Gamma^{\text{EOp}}(f,S)) \geq 8\sqrt{2\frac{d\log(\frac{2eNm}{d}) + \log(\frac{24}{\delta})}{NP^2}}\right) \leq \delta \tag{39}$$

whenever:

$$1 > 8\sqrt{2\frac{d\log(\frac{2eNm}{d}) + \log(\frac{24}{\delta})}{NP^2}} \geq 4\sqrt{\frac{2\log(12)}{NP^2}}$$

It is easy to see that the right inequality holds whenever $d \geq 1$, $2Nm \geq d$ and $\delta < 1$. In addition, inequality (39) trivially holds if the left inequality is not fulfilled. Hence the result follows. □

B. Details of Experimental Setup

Here we present details on the data preprocessing for the two dataset we use in our experiments, in particular the construction of the feature embeddings $\phi : \mathcal{Q} \times \mathcal{D} \to \mathbb{R}^D$.

TREC[2]**.** Inspired by the learning to rank approach for the TREC track of (Bonart, 2019), we pre-compute 9-dimensional embeddings of every query-paper pair by using the following hardcrafted features:

- the BM25 score of the query with the title, abstract, authors, topics and publication venue of the paper (5 values),
- the number of in- and out-citations (2 values),
- the publication year of the paper (1 value)
- the character length of the query (1 value).

Each feature is normalized by substracting the mean of the feature over the dataset and dividing by its standard deviation.

MSMARCO[3]**.** We use pretrained 768-dimensional BERT feature embeddings (Devlin et al., 2019) for representing the query-passage pairs. Specifically, we follow the embedding procedure described in (Nogueira & Cho, 2019; Han et al., 2020), where each query-passage pair is represented as the following token sequence:

$$[CLS] \text{ query text } [SEP] \text{ passage text } [CLS]$$

This sequence is then processed through a pre-trained BERT model[4] from Tensorflow Hub (Abadi et al., 2015), with maximum sequence length set to 200, and the hidden units of the first $[CLS]$ token are used as a representation of the query-passage pair.

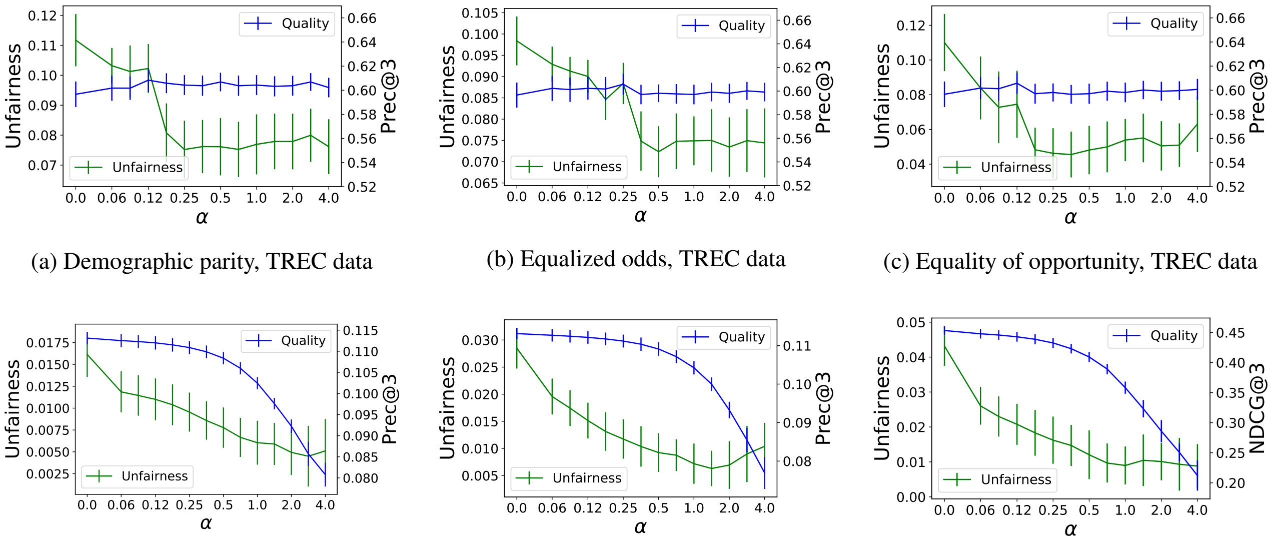
C. Further experimental results

We report on multiple additional metrics and experiments on the TREC and MSMARCO data, that were deferred to the supplementary material for space reasons. We also present more plots such as those in Figure 1, but for all values of $k = 1, 2, \ldots, 5$ and all splits into protected groups.

[2] https://fair-trec.github.io/2019/index.html

[3] https://microsoft.github.io/msmarco/

[4] https://tfhub.dev/tensorflow/bert_en_uncased_L-12_H-768_A-12/1



(a) Demographic parity, TREC data

(b) Equalized odds, TREC data

(c) Equality of opportunity, TREC data

(d) Demographic parity, MSMARCO

(e) Equalized odds, MSMARCO

(f) Equality of opportunity, MSMARCO

Figure 2: Results of our proposed method for learning fair rankers: unfairness (left y-axes) and ranking quality, measured through $\mathrm{P@k}$ (right y-axes) after training with different values of $\alpha \in [0, 4]$.

C.1. Additional metrics and experiments

Results with $\mathrm{P@k}$ We first present plots from the same experiments as in Figure 1, but with Precision@k as a metric for model performance. Specifically, Figure 2 shows the results when imposing different amounts of the three discussed fairness notions in typical settings for TREC ($t = 3$, $k = 3$; top row) and MSMARCO (*com*, $k = 3$; bottom row). We see a similar picture as with the NDCG metric, with no loss in precision for the TREC data and little to no effect for MSMARCO, for small to medium values of α.

Mean improvement of fairness over multiple α values We also present a robust analog of Table 1. Specifically, Table 2 reports *the average improvement in fairness over all values of α for which the NDCG was not affected significantly*, for both datasets and all experimental setups. The numbers confirm the observations from the main body of the paper that fairness can be increased substantially without affecting the model performance. Importantly, this holds on average over many values of α, suggesting that the observation is quite robust to the exact choice of the α parameter.

Enforcing fairness per query We also perform experiments for a variant of our algorithm that enforces the fairness notions for each query individually, rather than on average over all queries, similarly to (Singh & Joachims, 2017). In our framework this is achieved by regularizing with a separate term for every query in a batch and then averaging over the batch afterwards. We present the results in Table 3. We again show the maximal possible improvement of the fairness metric that is possible without a significant decrease in the NDCG. Note that we report the values of the original fairness violations

Table 2: Mean improvement in fairness possible without a significant decrease in NDCG, over multiple experimental setups.

TREC		k					average over k
		1	2	3	4	5	
equality of opportunity	$t = 3$	26%	39%	45%	27%	24%	32%
	$t = 4$	42%	36%	35%	41%	28%	36%
	$t = 5$	36%	33%	42%	38%	18%	33%
demographic parity	$t = 3$	27%	19%	24%	8%	10%	18%
	$t = 4$	51%	25%	32%	28%	23%	32%
	$t = 5$	44%	24%	55%	40%	39%	40%
equalized odds	$t = 3$	13%	12%	17%	9%	12%	12%
	$t = 4$	25%	17%	22%	25%	18%	21%
	$t = 5$	24%	11%	27%	27%	18%	21%
average over settings		32%	24%	33%	27%	21%	27%

MSMARCO		k					average over k
		1	2	3	4	5	
equality of opportunity	*com*	49%	39%	39%	36%	34%	39%
	ext	18%	11%	11%	10%	16%	13%
demographic parity	*com*	23%	26%	27%	31%	35%	28%
	ext	6%	9%	12%	13%	17%	12%
equalized odds	*com*	49%	41%	42%	40%	41%	43%
	ext	19%	17%	16%	17%	24%	19%
average over settings		27%	24%	25%	25%	28%	26%

measures (amortized over all queries), which is also supposed to be small, provided that the fairness on the per-query level is satisfied.

We see from Table 3 that although some improvement in fairness is possible with this method as well, the gains are much smaller and inconsistent between experiments. This suggests that averaging over multiple queries is important for optimization purposes, as the ranker has more flexibility to assign top-k positions fairly without sacrificing performance.

Table 3: Results of enforcing the fairness notions for every individual query separately. The table gives the maximal improvement in fairness possible without a significant decrease in NDCG (model performance), over multiple experimental setups.

TREC		k					average over k
		1	2	3	4	5	
equality of opportunity	$t = 3$	0%	0%	0%	0%	6%	1%
	$t = 4$	22%	21%	21%	40%	14%	24%
	$t = 5$	0%	21%	34%	29%	11%	19%
demographic parity	$t = 3$	0%	0%	8%	0%	3%	2%
	$t = 4$	10%	26%	24%	22%	15%	19%
	$t = 5$	33%	29%	43%	17%	15%	28%
equalized odds	$t = 3$	0%	0%	1%	0%	1%	0%
	$t = 4$	11%	27%	19%	20%	20%	20%
	$t = 5$	8%	17%	17%	18%	4%	13%
average over settings		9%	16%	19%	16%	10%	14%

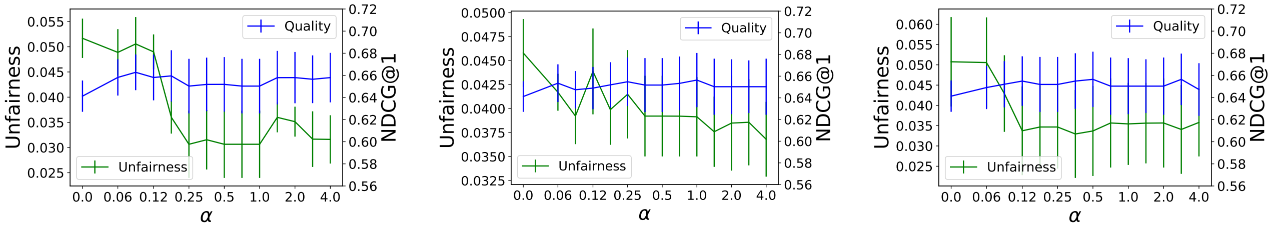
MSMARCO		k					average over k
		1	2	3	4	5	
equality of opportunity	*com*	29%	25%	22%	17%	16%	22%
	ext	8%	2%	1%	0%	2%	3%
	com	38%	36%	35%	34%	0%	29%
	ext	3%	3%	5%	5%	6%	4%
	com	16%	16%	14%	13%	12%	14%
	ext	6%	1%	1%	1%	2%	2%
average over settings		17%	14%	13%	12%	6%	12%

C.2. Plots for other values of k and other splits into protected groups

All plots below show NDCG@k and fairness, for the three fairness notions on every row.

C.2.1. TREC RESULTS

Different rows correspond to different values of k and t (the threshold for the i10 index).

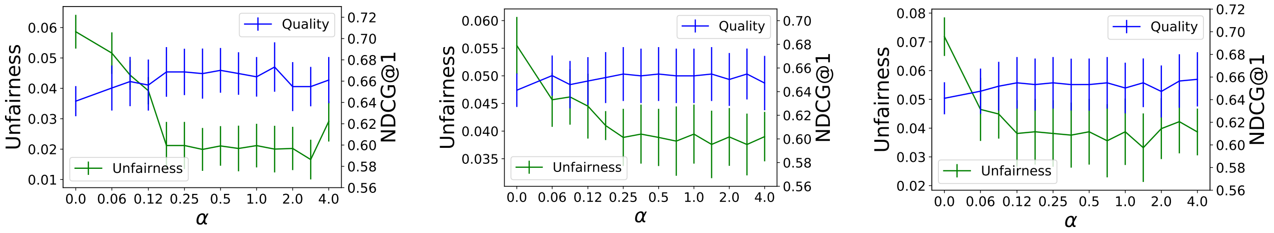


(a) Demographic parity, TREC data

(b) Equalized odds, TREC data

(c) Equality of opportunity, TREC data

Figure 3: $k = 1, t = 3$

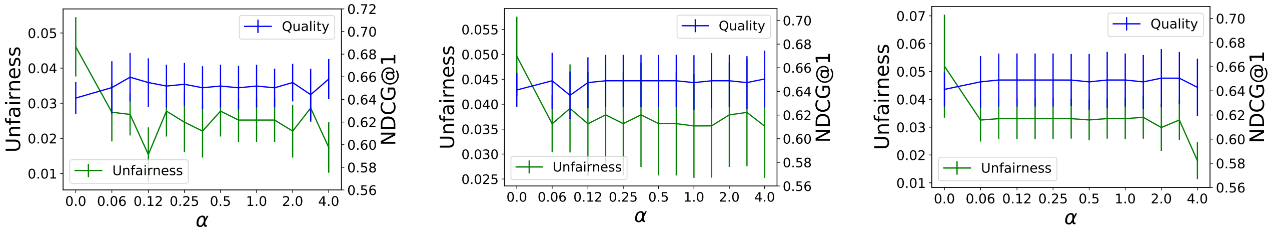


(a) Demographic parity, TREC data

(b) Equalized odds, TREC data

(c) Equality of opportunity, TREC data

Figure 4: $k = 1, t = 4$

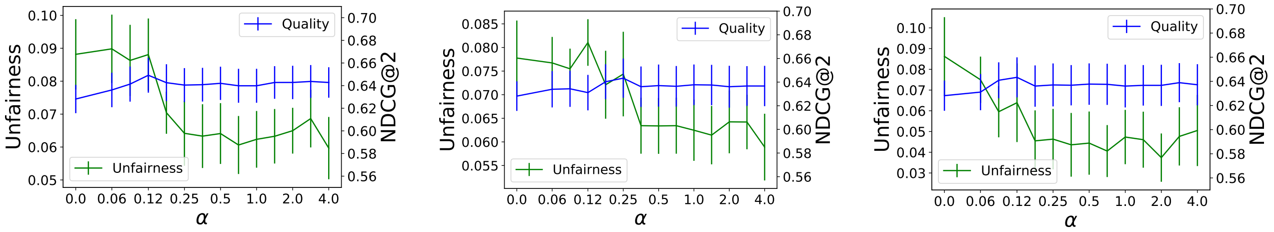


(a) Demographic parity, TREC data

(b) Equalized odds, TREC data

(c) Equality of opportunity, TREC data

Figure 5: $k = 1, t = 5$

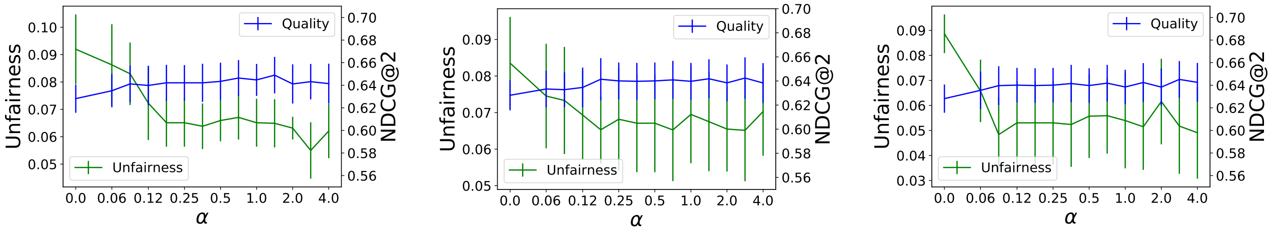


(a) Demographic parity, TREC data

(b) Equalized odds, TREC data

(c) Equality of opportunity, TREC data

Figure 6: $k = 2, t = 3$

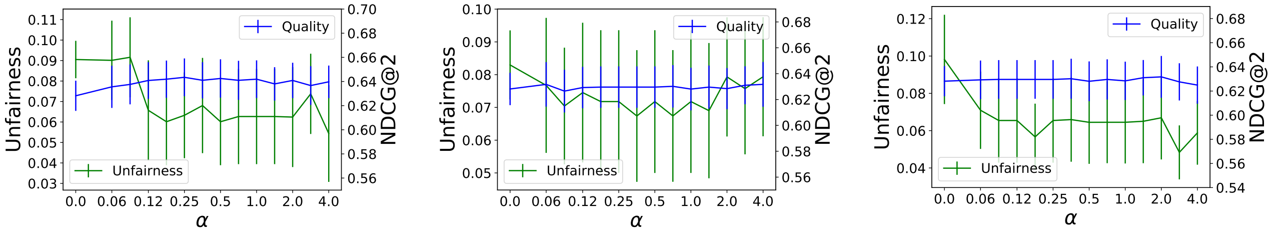


(a) Demographic parity, TREC data

(b) Equalized odds, TREC data

(c) Equality of opportunity, TREC data

Figure 7: $k = 2, t = 4$

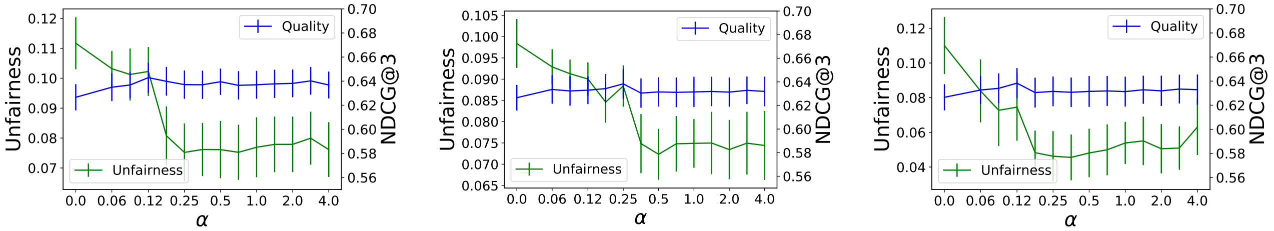


(a) Demographic parity, TREC data

(b) Equalized odds, TREC data

(c) Equality of opportunity, TREC data

Figure 8: $k = 2, t = 5$

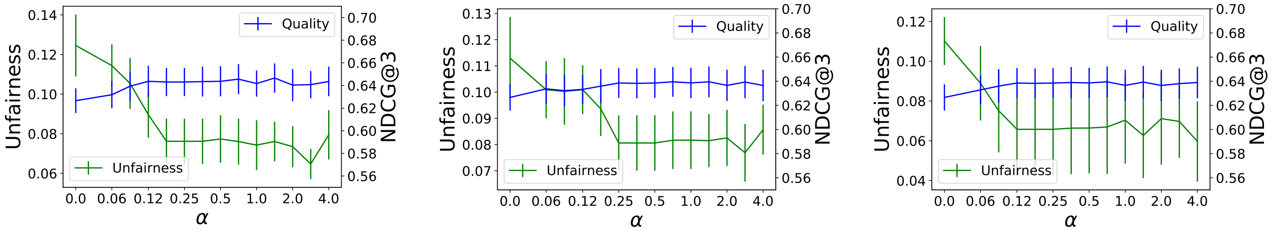


(a) Demographic parity, TREC data

(b) Equalized odds, TREC data

(c) Equality of opportunity, TREC data

Figure 9: $k = 3, t = 3$

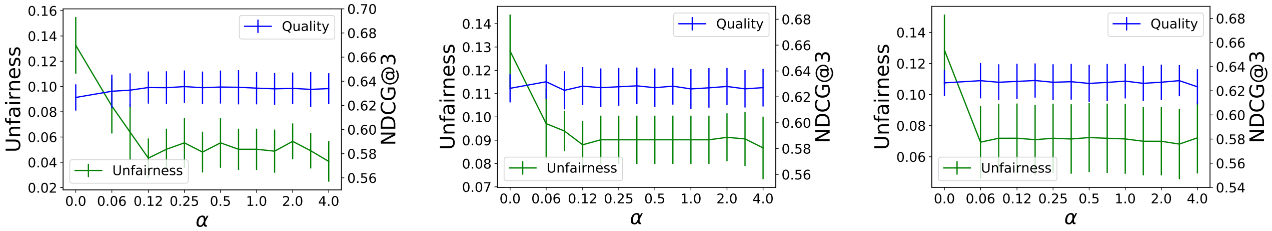


(a) Demographic parity, TREC data

(b) Equalized odds, TREC data

(c) Equality of opportunity, TREC data

Figure 10: $k = 3, t = 4$

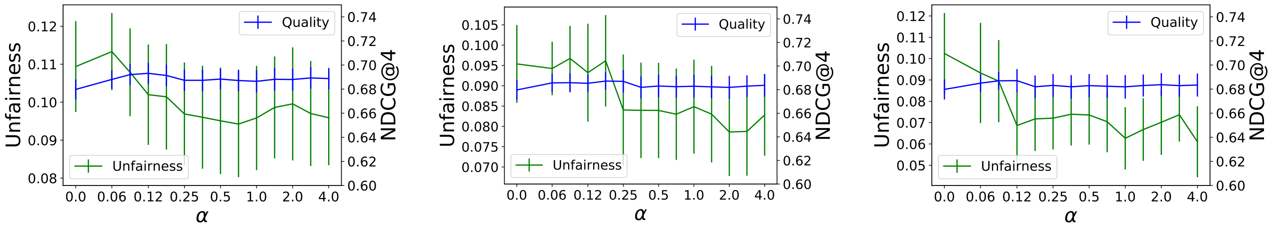


(a) Demographic parity, TREC data

(b) Equalized odds, TREC data

(c) Equality of opportunity, TREC data

Figure 11: $k = 3, t = 5$

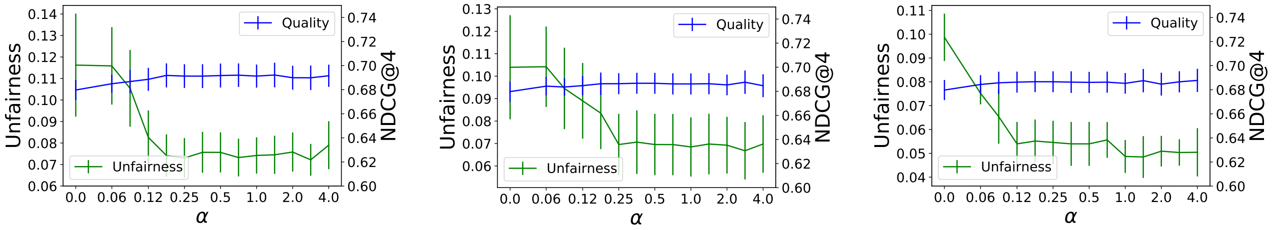


(a) Demographic parity, TREC data

(b) Equalized odds, TREC data

(c) Equality of opportunity, TREC data

Figure 12: $k = 4, t = 3$

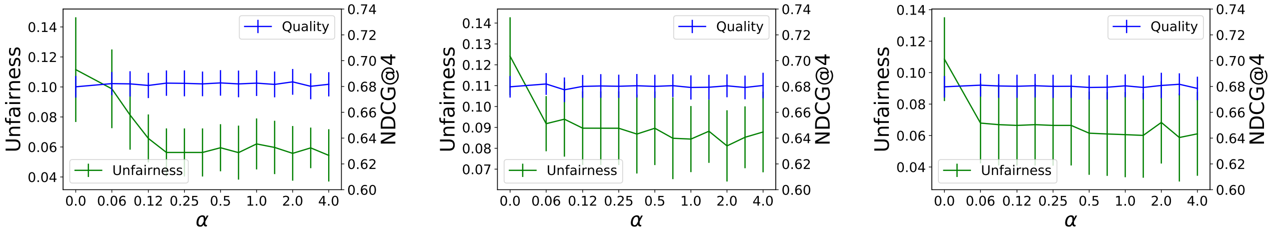


(a) Demographic parity, TREC data

(b) Equalized odds, TREC data

(c) Equality of opportunity, TREC data

Figure 13: $k = 4, t = 4$

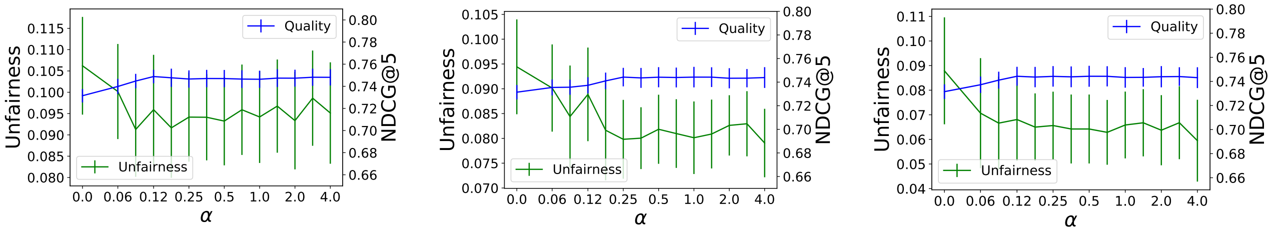


(a) Demographic parity, TREC data

(b) Equalized odds, TREC data

(c) Equality of opportunity, TREC data

Figure 14: $k = 4, t = 5$

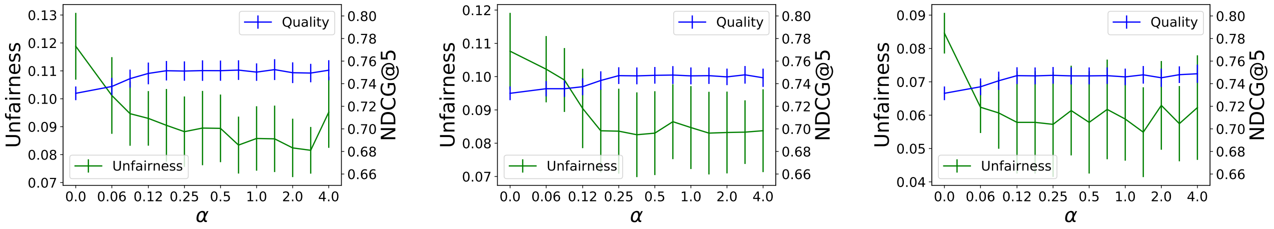


(a) Demographic parity, TREC data

(b) Equalized odds, TREC data

(c) Equality of opportunity, TREC data

Figure 15: $k = 5, t = 3$

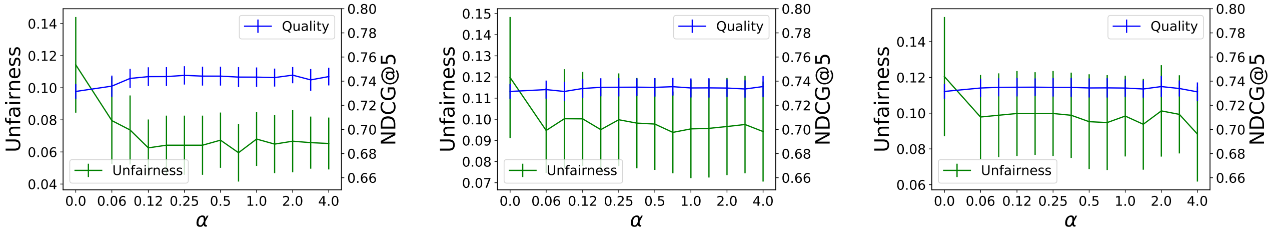


(a) Demographic parity, TREC data

(b) Equalized odds, TREC data

(c) Equality of opportunity, TREC data

Figure 16: $k = 5, t = 4$



(a) Demographic parity, TREC data

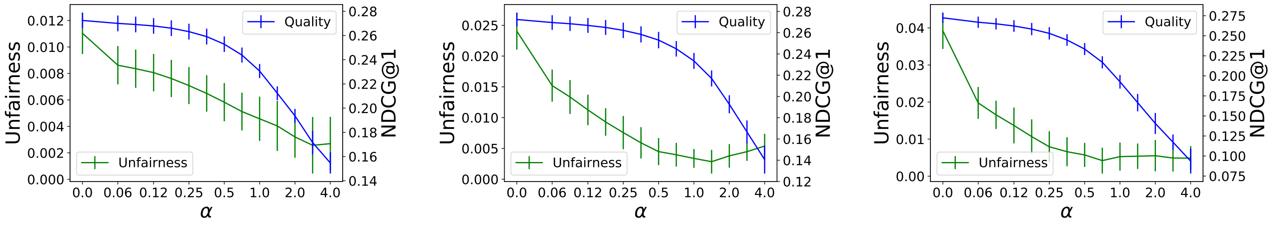
(b) Equalized odds, TREC data

(c) Equality of opportunity, TREC data

Figure 17: $k = 5, t = 5$

C.2.2. MSMARCO RESULTS

Different rows correspond to different values of k and the two different splits into protected groups (*com* and *ext*).

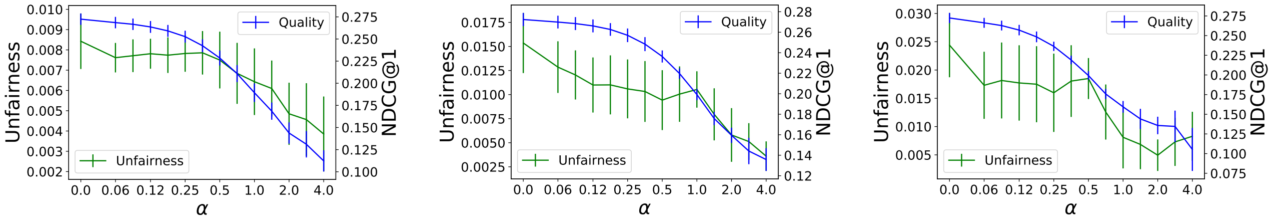


(a) Demographic parity, MSMARCO

(b) Equalized odds, MSMARCO

(c) Equality of opportunity, MSMARCO

Figure 18: $k = 1$, *com*

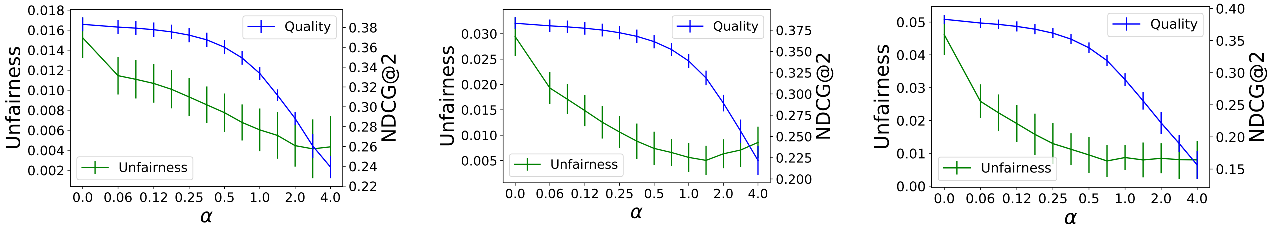


(a) Demographic parity, MSMARCO

(b) Equalized odds, MSMARCO

(c) Equality of opportunity, MSMARCO

Figure 19: $k = 1$, *ext*

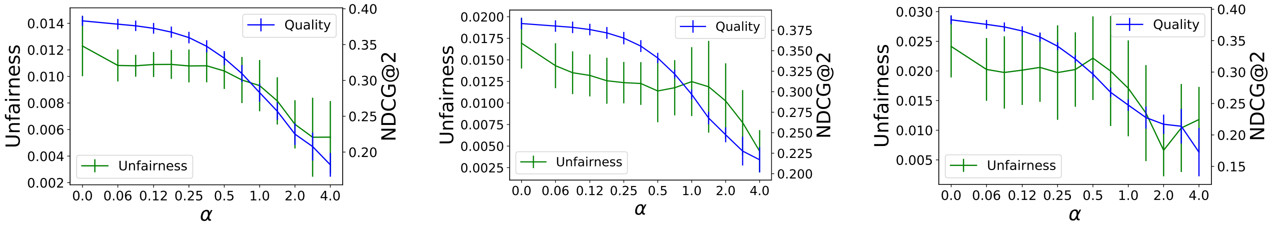


(a) Demographic parity, MSMARCO

(b) Equalized odds, MSMARCO

(c) Equality of opportunity, MSMARCO

Figure 20: $k = 2$, *com*

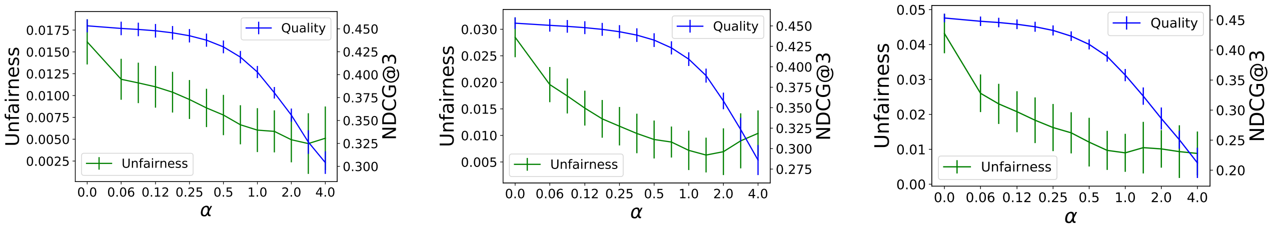


(a) Demographic parity, MSMARCO

(b) Equalized odds, MSMARCO

(c) Equality of opportunity, MSMARCO

Figure 21: $k = 2$, *ext*

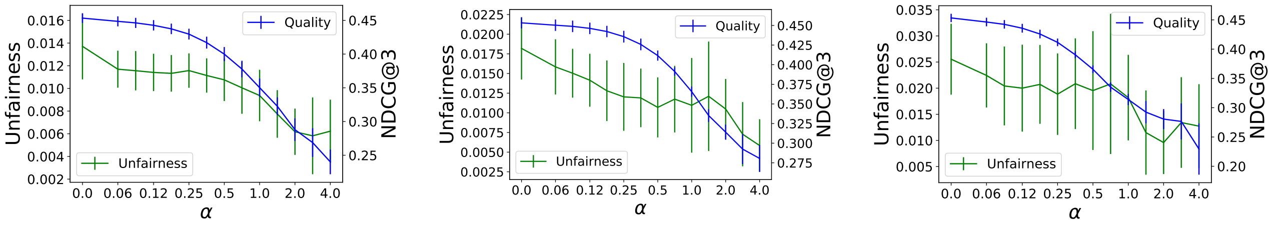


(a) Demographic parity, MSMARCO

(b) Equalized odds, MSMARCO

(c) Equality of opportunity, MSMARCO

Figure 22: $k = 3$, *com*

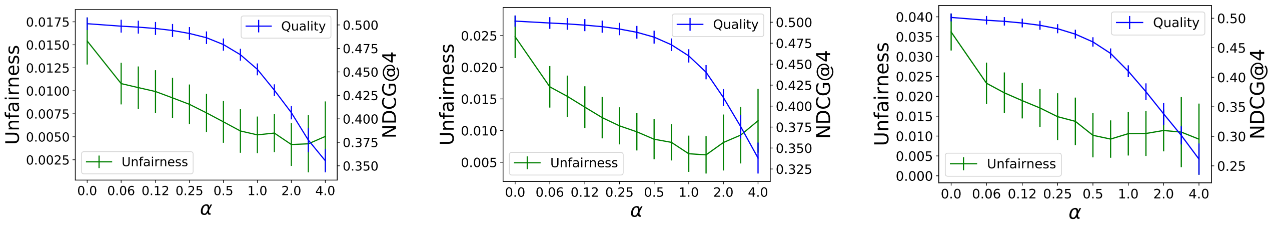


(a) Demographic parity, MSMARCO

(b) Equalized odds, MSMARCO

(c) Equality of opportunity, MSMARCO

Figure 23: $k = 3$, *ext*

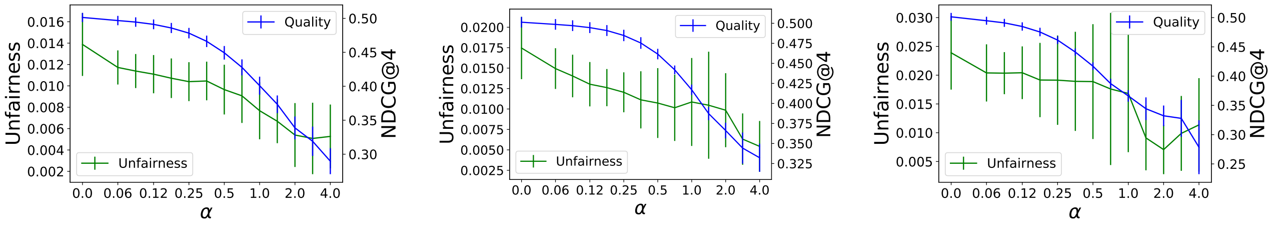


(a) Demographic parity, MSMARCO

(b) Equalized odds, MSMARCO

(c) Equality of opportunity, MSMARCO

Figure 24: $k = 4$, *com*

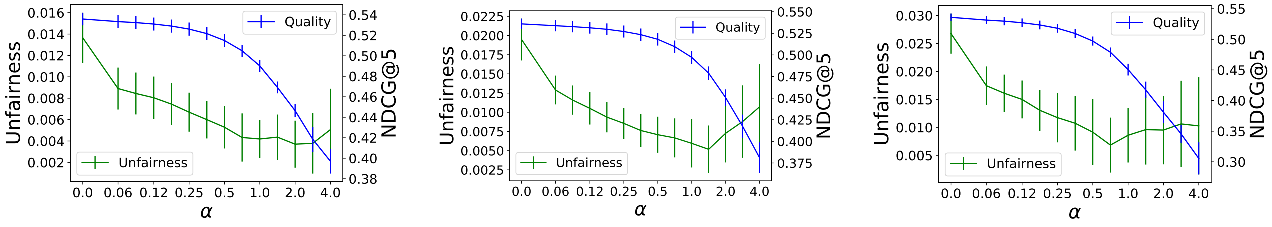


(a) Demographic parity, MSMARCO

(b) Equalized odds, MSMARCO

(c) Equality of opportunity, MSMARCO

Figure 25: $k = 4$, *ext*

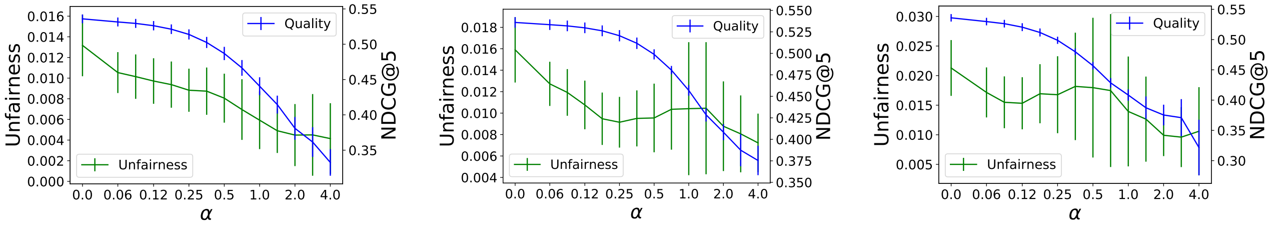


(a) Demographic parity, MSMARCO

(b) Equalized odds, MSMARCO

(c) Equality of opportunity, MSMARCO

Figure 26: $k = 5$, *com*



(a) Demographic parity, MSMARCO

(b) Equalized odds, MSMARCO

(c) Equality of opportunity, MSMARCO

Figure 27: $k = 5$, *ext*