Fair Pairwise Learning to Rank

Fair pairwise learning to rank

Mattia Cerrato*§, Marius Köppel†§, Alexander Segner†, Roberto Esposito*, Stefan Kramer†
* *Università di Torino*
Via Pessinetto 12
Torino, Italy
†*Johannes Gutenberg-Universität Mainz*
Saarstraße 21
Mainz, Germany

***Abstract*—Ranking algorithms based on *Neural Networks* have been a topic of recent research. Ranking is employed in everyday applications like product recommendations, search results, or even in finding good candidates for hiring. However, Neural Networks are mostly opaque tools, and it is hard to evaluate why a specific candidate, for instance, was not considered. Therefore, for neural-based ranking methods to be trustworthy it is crucial to guarantee that the outcome is fair and that the decisions are not discriminating people according to sensitive attributes such as gender, sexual orientation, or ethnicity.**

In this work we present a family of *fair pairwise learning to rank* approaches based on Neural Networks, which are able to produce balanced outcomes for underprivileged groups and, at the same time, *build fair representations of data*, i.e. new vectors having no correlation with regard to a sensitive attribute. We compare our approaches to recent work dealing with fair ranking and evaluate them using both relevance and fairness metrics. Our results show that the introduced fair pairwise ranking methods compare favorably to other methods when considering the fairness/relevance trade-off.

***Index Terms*—fairness, neural networks, ranking**

I. Introduction

Information retrieval plays a prominent role in most machine learning applications. Retrieving relevant data is a critical task with many applications in research and industry. One of the problems tackled by information retrieval is *learning to rank* [11], [27], where documents[1] are sorted according to their relevance to a query. This kind of model may also be employed in tasks which have a clear and immediate impact on human well-being, such as ranking candidates for admission to higher education or giving out scholarships. Another example is the COMPAS software, which has been developed to assist court judges in deciding whether an individual is at risk of reoffending and, ultimately, whether the person could be released on parole. It has been shown [2] that the software can be biased against black people, in the way that they are consistently mis-assigned higher risk scores.[2]

In this situation, it is unclear whether ranking machine learning systems can be considered trustworthy, as models which rely on statistics extracted from biased data can perpetrate and justify further discriminating behavior. While the definitions of "bias" and "fairness" are still a topic of active academic discussion, it is not a newly found concern, in neither machine learning nor computer science, where discussions on the matter date back to the nineties [17]. Lately, however, there has been a growing call from both regulatory institutions [12] and the general public [2] for automated decision processes to be fair and transparent. This problem is exacerbated by the trend in employing opaque deep neural models in many machine learning tasks, including ranking [27]. Making sense (in a human-intelligible way) of the decisions undertaken by automated algorithms is the most straightforward path to solve the fairness problem since a good explanation would allow the end user to exclude unfair reasons from the motivations underlying the decision. Model explanation, despite being a relevant and very active line of research [1], [7], is far from being a solved problem, especially for large complex models with millions of parameters. Our approach is instead to condition the learning of the system in such a way that it guarantees that no sensitive information is used to take the final decision. To do so, we employ two different mechanisms which *predict* and *rank* the sensitive attribute based on the internal representation of the data built by the *main* network. The output of these networks is used to penalize the main network every time the sensitive attribute is correctly predicted or ranked. This can be done by inverting the gradient's sign when it is backpropagated in the main network [20], [38]. We demonstrate that these approaches, when combined with further bias-reduction mechanisms [9], can obtain fair, trustworthy models and representations. The contributions of the paper are thus as follows:

(i) We introduce a new family of *fair pairwise ranking methods* that builds on the *pairwise ranking model* DirectRanker [25] and investigate the usefulness of the Gradient Reversal Layer by Ganin et al. [20]. More specifically, we introduce two different mechanisms to obtain ranking models that are also fair.
(ii) We show how our models are able to output results which are both high in relevance and unbiased on standard

§These authors contributed equally.

[1]We use the terms *documents* and *instances* as synonyms in the remainder of the paper.

[2]The more general claim that the software is "unfair" has been challenged, see the report from Northpointe, the company that developed COMPAS [13], a discussion of the applicability of the "disparate impact" fairness criterion [10], and a discussion on transparency and the role of proprietary software in the courtroom [34].

datasets that are widely employed in the fairness literature.

(iii) We compare our methods with a fair *listwise* learning to rank approach called *DELTR* [44], a fair classifier [9], [38], and with a constraint optimized *pairwise* ranker [31] and show that they are able to obtain rankings which are more fair and just as relevant or better.

The paper is structured as follows. We discuss the underlying models related to our approach in Section II and outline the differences to other fair ranking methodologies. The models and their properties are discussed in Section III, before we lay out details of the experimental setup in Section IV. Following the discussion of results in Section V, we draw our conclusions in Section VI.

II. Related Work

In this section, we give an overview of related work, starting with learning to rank approaches, followed by adversarial approaches to address fairness in classification, and recent work on fair ranking.

A. Learning to rank

Models that address the *learning to rank* problem, in which a list of n documents needs to be sorted based on relevance to a query, fall into three broad categories according to whether the objective function is computed by considering one, two or the whole list of documents at a time during training. The first approach is called *pointwise* and is analogous to classifying each document [11], [26] in the sense that instead of comparing documents in a list, a score is predicted on each query-document pair, indicating the single document's relevance to the query. In the *pairwise* approach a model tries to determine the more relevant document out of two for the given query [4], [18]. The last approach is called *listwise*, in which the whole list is used to compute the cost during training [6], [39].

It is possible to extend many classification algorithms to the ranking problem, such as decision trees [18], support vector machines [5], artificial neural networks [6] and ensemble boosting [37].

While the listwise approach allows for direct optimization of the desired listwise metrics, more recently it has been shown that a generalization [25] of the *pairwise* learning approach RankNet [4] outperforms numerous state-of-the-art methods on NDCG and MAP while requiring much shorter training times. This learning algorithm has been proven to be able to learn a total quasiorder on a wide variety of feature spaces by having the requirements for such an order inherently built into its network architecture using a siamese structure. Our methodology builds on such a ranker by adding mechanisms which are useful in making the resulting model also fair.

B. Fairness

Defining fairness in machine learning has attracted considerable attention. Two possible broad categories in which different definitions fall into are *individual* and *group* fairness. Individual fairness is defined as treating similar individuals similarly [16]. In this paper we deal with group fairness, where the focus is instead shifted on group-wise definitions and metrics. In classification tasks, possible definitions include disparate impact, disparate treatment, and disparate mistreatment [42]. A model displays *disparate impact* when it assigns positive outcomes with different rates to individuals belonging to different groups. *Disparate treatment* can be observed when a classifier provides different outputs for people belonging to different groups who are otherwise similar. Lastly, *disparate mistreatment* refers to the situation where a decision system displays different error rates for individuals belonging to different groups, as it was observed for the COMPAS software [2]. These definitions, in practice, are applied by evaluating for instance the difference in positive outcomes for individuals belonging to protected and non-protected groups. An individual belongs to a protected group when it is thought or known that data contains discriminatory or biased information about the same group. In the aforementioned case of predicting future re-offenders, the concern is that the historically higher rate of incarceration for black individuals [35] can produce biased data which will lead to biased algorithms when employed as training information.

1) Fair classification: In the context of fair classification, we consider a dataset $D = \{(x_i, s_i, y_i), i \in \{1...N\}\}$ where $x_i \in \mathbb{R}^m$ are vectors describing the non-sensitive attributes of the documents, $s_i \in \mathbb{R}^n$ are vectors describing sensitive attributes (for instance, ethnicity or gender), and y_i represents the target value of each instance (for instance, getting a loan or being admitted to higher education programs). Such a setup has been investigated extensively. Some approaches propose preprocessing of the data [22] so that positive outcomes of $y = 1$ in the training set are balanced between different groups, which are represented as different values of a sensitive attribute s. Other strategies are regularization-based and insert "fairness" into the objective function in various ways. Algorithms that can be adapted to fair classification include logistic regression [23], probabilistic models [21], and discriminative clustering models [45]. More relevant to our approach is the neural-based Domain Adversarial classifier [20]. This model is able to learn a shared representation of data coming from different domains (datasets) by "un-learning" information about specific domains. Since then, the framework has been adapted to fair classification [29], [38], where it is instead information about a sensitive attribute that is removed. However, it has been shown [9] that the base strategy of employing Gradient Reversal is not enough to guarantee that the features extracted by the network will be debiased, i.e., decorrelated with respect to the sensitive attribute. We describe how these techniques can be employed in fair ranking in Section III.

2) Fair Ranking: In fair ranking one wants to find a quasiorder of documents according to their relevance while guaranteeing some notion of fairness with respect to a sensitive attribute. Let us consider a dataset $D = \{(q_i, x_i, s_i, y_i), i \in \{1...N\}\}$ where q_i are the queries, $x_i \in \mathbb{R}^m$ are vectors describing the non-sensitive attributes of the documents, $s_i \in \mathbb{R}^n$ are vectors describing sensitive attributes, and y_i represents each document's relevance for the query. The main

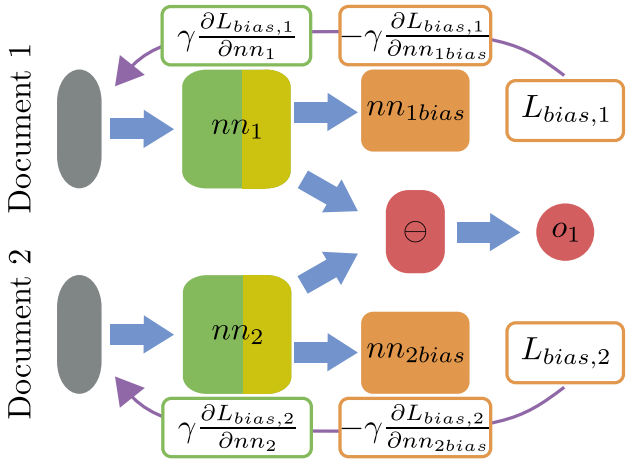
motivation for exploring the task of ranking fairly is ensuring that individuals belonging to protected groups are not relegated to lower ranking positions [41] because of past or present human biases which may be reflected in training data. A number of different metrics have been introduced for the purpose of evaluating ranked outputs for their group fairness. Ke and Stoyanovich [40] propose metrics which take into account the proportion of individuals coming from underpriviledged groups belonging in the top-i positions of the model output (also see Section IV-A1). As protected group membership should not influence an individual's position in the ranked output, the authors' metrics can help evaluate the disparate impact of a ranking model. Singh and Joachims [33] have instead focused on *average exposure*, which is defined as the average probabilities of all individuals belonging to a specific group to be ranked at the top of the list. More recently, Narasimhan et al. [31] have argued for ranking fairness to be measured, in continuity with the disparate mistreatment concept in classification, as the difference in rank accuracy between different groups (Section IV-A2). Learning frameworks in fair ranking have so far been focused on post-processing methods [43], [8], [33]. A post-processing method first ranks individuals by only taking query relevance into account, and fairness is only enforced by re-ranking. Some authors [44] have however raised concerns about the legal admissibility of these methods. Regularization techniques instead optimize an objective which takes into consideration both relevance and fairness. Zehlike et al. [44] propose a listwise ranking method with a term encouraging a balanced *average exposure* for different groups [33]. The methodology can be thought of as a fair extension to a listwise ranking model. Another neural-based, fair pairwise ranking model [31] generalizes to various definitions of fairness by casting them as constrained optimization problems. In contrast, our model actively removes sensitive information from the training vectors and is able to generalize well to different fairness ranking measures.

III. Model Description

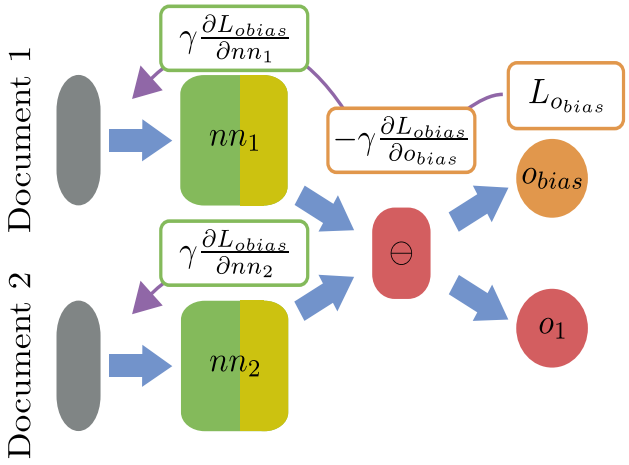
In the following we describe our contribution. In the first two sections we discuss the general elements of our methodology, the DirectRanker [25] model and the Gradient Reversal layer [20]. Given these preliminaries, we will explain two new methodologies that are able to learn relevant and fair rankings (Sections III-C and III-D). We will also investigate the combination of these methodologies with a noise conditioning layer (Section III-E).

A. DirectRanker

The DirectRanker model [25] has been introduced as a generalization of the RankNet architecture [4]. This model is constrained to learn a total quasiorder on the feature space. As such, the architecture of the model includes a pairwise ranking function which is reflexive, antisymmetric and transitive by construction. Reflexivity is enforced by employing two networks which process different documents whilst sharing their parameters and architecture. These networks are called



(a) Adversarial Fair DirectRanker



(b) Fair Flipped DirectRanker

Fig. 1. Schema of the Adversarial (top) and Fair Flipped (bottom) DirectRanker models. nn_1 and nn_2 can be arbitrary networks (or other function approximators) as long as they give the same output for the same inputs, meaning that they share their weights. In both models, the output neuron o_1 predicts the difference between the relevance labels while the bias is zero and the activation can be any antisymmetric, sign-conserving function. In the adversarial model, The $nn_{1_{bias}}$ and $nn_{2_{bias}}$ parts predict the sensitive attributes of the two documents. In the backwards step, the gradients are flipped when backpropagated into the feature extraction layers. In the fair flipped model, the o_{bias} neuron predicts the difference of the sensitive attributes of the two documents while also inverting the gradients when backpropagating into the feature part (nn_1 and nn_2). The yellow part of nn_1 and nn_2 shows the last layer, which is used for extracting the unbiased representations.

nn_1 and nn_2 in Figs. 1(a) and 1(b). Antisymmetry is granted by constraining the architecture to have sign-conserving neural activation and no bias in the output neuron (o_1), and having it only depend on the difference of the two previously mentioned networks, as an antisymmetric input [32]. The authors also detail how to prove that this model is forced to have a transitive ranking function. This model has been shown to be as good as listwise rankers, while also being able to be trained in a fraction of the time. Due to these favorable properties, we will show how to derive a fair pairwise ranking scheme on the basis of the DirectRanker. We refer the reader to Köppel et al. [25] for full details of the model.

B. Gradient Reversal

The concept of a Gradient Reversal layer has been first introduced in domain adaptation [20] and since then generalized to fair classification [38], [9], [29]. Domain adaptation is the multi-task learning setup where one has access to labeled examples $(x_i, y_i)_{i=1}^n$ on a *source* domain and to unlabeled examples $x_{j=i+1}^m$ on a *target* domain. One possible approach to exploit knowledge from both domains is to learn a shared feature space where learning algorithms are unable to identify whether an example was sampled from the source or target distribution [3]. In essence, this approach proposes to learn representations of the data which are domain-invariant. In this setting, Ganin *et al.* [20] contributed a neural network model which is trained to optimize two training objectives at the same time, one expressing the risk on the target domain and another representing the distance between the source and target domains. In their work, the two objectives are optimized by one sub-network each. These sub-networks are optimized to predict the labels and the domain, respectively, from the features extracted by a shared network. When backpropagating the domain gradients into the network, the gradient's sign is, however, inverted. Ganin *et al.* show that this learning scheme is able to find a saddle point equilibrium between the two aforementioned objectives. In fair classification, learning representations which are invariant to the sensitive attribute is helpful in removing the complex, non-linear correlations which can still identify a person's gender or race even when removing sensitive features [28], [38]. In this work, we employ a Gradient Reversal layer in two separate ways. In the first one (Section III-C), we add sub-networks that predict the sensitive attributes from the representations that the DirectRanker model is learning. Another option (Section III-D) is to instead add another output neuron to the DirectRanker model and train its parameters to predict whether the two input documents have the same or a different value for the sensitive attribute. In both cases, we invert the gradients when back-propagating in the main feature extraction networks.

C. Fair Adversarial DirectRanker

Following the strategy outlined in the previous section, the twin feature extraction networks of the DirectRanker can be constrained for fairness by using two auxiliary networks $nn_{1_{bias}}$ and $nn_{2_{bias}}$, which predict the sensitive attribute for a given individual/training sample. In contrast to the networks nn_1 and nn_2 used for the ranking part, $nn_{1_{bias}}$ and $nn_{2_{bias}}$ do not need shared parameters. The model architecture is shown in Fig. 1(a).

As the gradient's sign is inverted when backpropagating into the feature extraction layers, this has the effect of removing information about the sensitive attribute in a feature extraction network [9], [29], [38], therefore encouraging the representation to be fair.

Since this approach requires a loss function for both the ranking and the prediction of the sensitive attributes, we construct our loss in the following way:

$$\begin{aligned} L(\Delta y, x_1, x_2, s_1, s_2) = L_{\text{rank}}(\Delta y, x_1, x_2) \\ + \gamma \sum_{i=1}^{2} L_{\text{bias},i}(s_i, x_i), \end{aligned} \tag{1}$$

where the γ hyperparameter regulates the relevance-fairness trade-off. This parameter therefore provides an option to balance the importance of removing information about s versus accurately predicting the relevance of the documents with respect to the ground truth. As a ranking loss we keep the choice employed in the original DirectRanker paper, namely

$$L_{\text{rank}}(\Delta y, x_1, x_2) = (\Delta y - o_1(x_1, x_2))^2 \tag{2}$$

as a loss function for the prediction of the relative ranking of instances (x_1, s_1, y_1, q) and (x_2, s_2, y_2, q) with $\Delta y = y_1 - y_2$. As for L_{bias}, any classification loss can be employed, such as

$$\begin{aligned} L_{\text{bias},i}(s, x) = -s \log(nn_{i\,bias}(x)) \\ - (1 - s) \log(1 - nn_{i\,bias}(x)), \end{aligned} \tag{3}$$

when s is binary. In the following, we will refer to this model as the Fair Adversarial DirectRanker (ADV DR).

D. Fair Flipped DirectRanker

Here we present another possible fair ranker based on DirectRanker: the Fair Flipped DirectRanker (see Fig. 1(b)). In contrast to the Fair Adversarial Direct Ranker, here we do not try to directly predict the sensible attribute s. Instead, we add another output neuron o_{bias} to the DirectRanker model (having the same properties as the ranking neuron o_1, namely a sign-conserving and antisymmetric activation function) that tries to predict whether the two input documents (x_i, s_i, y_i, q), $i = 1, 2$ have the same sensitive attribute s_i. We train this neuron using a quadratic loss function similar to the ranking loss:

$$L_{o_{bias}}(\Delta s, x_1, x_2) = (\Delta s - o_{bias}(x_1, x_2))^2 \tag{4}$$

with $\Delta s = s_1 - s_2$. As in the previous strategy, the gradient information is inverted when backpropagated through the feature extraction layers. The model is therefore optimized to be agnostic to the difference between the sensitive attribute of two input documents.

As for the ADV DR, the complete loss

$$\begin{aligned} L(\Delta y, \Delta s, x_1, x_2) = L_{\text{rank}}(\Delta y, x_1, x_2) \\ + \gamma L_{o_{bias}}(\Delta s, x_1, x_2) \end{aligned} \tag{5}$$

is a weighted combination of the ranking loss and the fairness loss. We keep the same ranking loss as for ADV DR, as first considered for the original DirectRanker architecture. In the following, we will refer to this model as FFDR for short.

TABLE I
OVERVIEW OF THE HYPERPARAMETERS INCLUDED IN OUR GRID SEARCH

parameter	values
activation function	tanh
#layers of $nn_{1/2}$	2
#neurons per layer in $nn_{1/2}$	5, 10, 20 ... 100
#layers of $nn_{bias_1/bias_2}$	2
#neurons per layer in $nn_{bias_1/bias_2}$	2, 10, 20 ... 50
training steps	0.5k, 1k, 1.5k ... 10k
γ	0.1, 1, 2, 5, 10, 100

E. Noise Conditioning Layer

As shown in related work [9], a noise conditioning layer can be used for simplifying the task of decorrelating the input features with respect to s. This layer can be added to any differentiable model and has been shown to aid neural networks in obtaining fair representations that contain little to no information about s. Noise conditioning layers can be inserted at any point in a network. Assuming that layer i computed the activations $\left(a_k^i\right)_{k=1}^n$, a noise layer at level $i+1$ computes the following:

$$a^{i+1} = a^i \odot w_1 + \eta \odot w_2, \tag{6}$$

where $\odot$ is the Hadamard product, w_1 and w_2 are learnable weight vectors, and η is a noise which is providing random vectors $\in \mathbb{R}^n$. As the feature extraction part of the DirectRanker can be any function, this can be easily added to the model. Note that the issue of this method is that if one would sample the noise at each forward pass of the network, the loss function would then not be functional univalent, as different outputs could be computed from the same input depending on the noise value. We follow here the same approach as before [9], meaning that we only sample once at the start of the learning process.

IV. EXPERIMENTAL SETUP

In the following we evaluate our models on ranking datasets commonly used in the fairness literature. To encourage fairness in the model's decisions, we employ two different mechanisms, which we evaluate separately for relevance and fairness by using standard metrics. We also show results for models which include both the mechanisms. For the relevance task, we use the commonly used nDCG@k and the AUROC, referred to here as AUC. The fairness of the models is evaluated via the rND metric (Section IV-A1) and explicitly by the accuracy obtained when predicting the sensitive attribute from the representations they learn. We also report the group-dependent pairwise accuracies (Section IV-A2), which have been recently developed [31]. We explored a number of hyperparameter combinations, which we report in Table I. Our evaluations are twofold, as we assess both the ranking models and their *extracted representations*, i.e. the features learned by the feature extraction layers. This can be done by training a supervised model on the aforementioned representations. This is a standard way to evaluate fair classifiers [28], [45], [9] and to understand how much information about s has been removed by the model, as high values for fairness can also be achieved by a weak ranker taking random or quasi-random decisions from biased data. We transpose this setup to the fair ranking paradigm by training both linear and non-linear classifiers and rankers (linear regression and random forest models, respectively). We then evaluate their invariance to the sensitive attribute and the fairness of the ranked outputs. Furthermore, this experimental setup can also simulate a "separations of concerns" regulatory scenario [29]. In this setting, users intending to train models which directly impact individuals only have access to representations where information about sensitive data was purged. This enables any downstream model to be fair by design, but poses the additional challenge of having to remove information from the data.

A. Evaluation Metrics

1) rND: To evaluate group fairness in the whole output list, we employ the rND metric introduced by Ke and Stoyanovitch [40]. The metric is defined as follows:

$$\text{rND} = \frac{1}{Z} \sum_{i \in \{10,20,\dots\}}^{N} \frac{1}{log_2 i} \mid \frac{\mid S^+_{1\dots i} \mid}{i} - \frac{\mid S^+ \mid}{N} \mid . \tag{7}$$

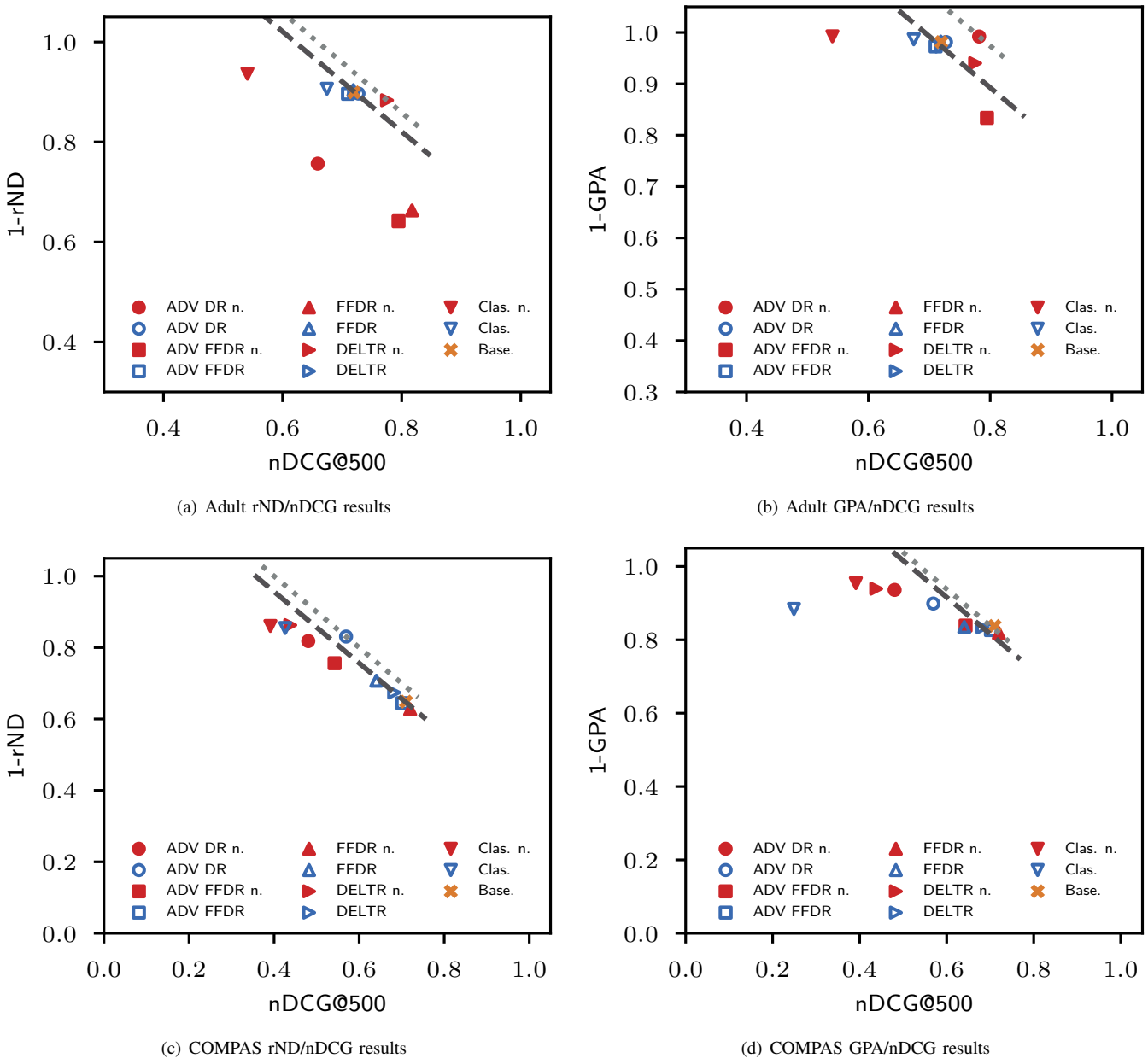
This metric computes the difference between the proportion of protected individuals in the top-i documents ($\frac{S^+_{1..i}}{i}$) and in the overall population ($\frac{S^+}{N}$). Z is a normalization factor which is defined as the maximum possible value of the metric. We evaluated this factor by computing the same metric over a dummy list in which all protected individuals can be found at the very end of the list. As argued by Ke and Stoyanovitch, this metric can be seen as a generalization of the disparate impact/statistical parity concept in fair classification. It warrants mentioning that over-representation (higher rate than the population proportion) of protected individuals at the top of the list is also penalized by the metric.

2) Group-dependent Pairwise Accuracy: Let $G_1, ..., G_K$ be a set of K protected groups such that every document inside the dataset D belongs to one of these groups. The *group-dependent pairwise accuracy* [31] $A_{G_i>G_j}$ is then defined as the accuracy of a ranker on documents which are labeled more relevant belonging to group G_i and documents labeled less relevant belonging to group G_j. Since a fair ranker should not discriminate against protected groups, the difference $|A_{G_i>G_j} - A_{G_j>G_i}|$ should be close to zero. In the following, we call the Group-dependent Pairwise Accuracy GPA.

B. Datasets

Our experimental evaluation is focused on datasets commonly employed in the fairness literature, such as Adult [15], COMPAS [2], and Law Students [44]. We also employed the Wiki Talk Page Comments dataset to enable a comparison with a recently developed neural-based constrained optimization method [31].

The Adult dataset's ground truth represents whether an individual's annual salary is over 50K$ per year or not [24].



(a) Adult rND/nDCG results

(b) Adult GPA/nDCG results

(c) COMPAS rND/nDCG results

(d) COMPAS GPA/nDCG results

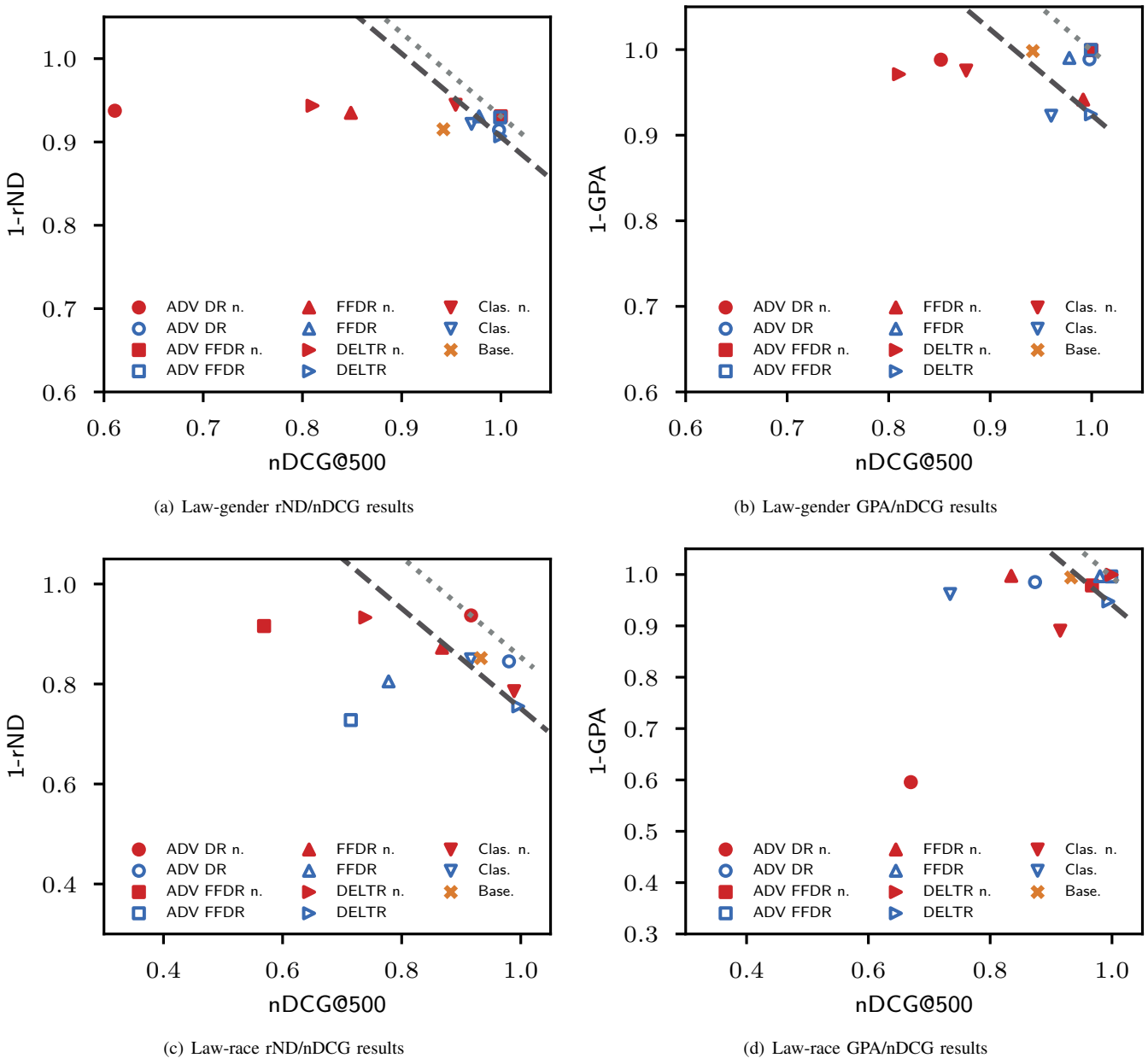
Fig. 2. Results for the *ADULT* and *COMPAS* datasets for the ranker outputs. Models marked with "n." such as ADV DR n. included a noise module in their architecture. The lines represent balanced fairness/relevance trade-offs. The dashed line represents all points with equal trade-off as the best performing comparison model, while the dotted line also takes into consideration the models we contribute.

It is commonly used in fair classification, since it is biased against gender [28], [45], [9].

The COMPAS dataset [2] has been released as part of an investigative journalism effort in tackling automated discrimination. The ground truth here is an individual's "risk score", which is supposed to be proportional to their probability of committing a crime in the near future. As in related work [42], [9], we focus on non-violent crimes and use the race attribute as the sensitive variable.

The Law Students dataset contains information relating to 21,792 US-based, first-year law students and was collected to the end of understanding whether the Law Students Admission Test in the US is biased against ethnic minorities [36]. As in related work [44], we subsampled 10% of the total samples while maintaining the distribution of gender and ethnicity, respectively. In this setting, it is possible to employ both variables as the sensitive attributes, which we did in two separate experimental setups, as done elsewhere. The task here is to sort students based on their predicted academic performance.

The Wiki Talk Page Comments dataset contains 127,820 Wikipedia comments which are labeled toxic or not toxic. A



(a) Law-gender rND/nDCG results

(b) Law-gender GPA/nDCG results

(c) Law-race rND/nDCG results

(d) Law-race GPA/nDCG results

Fig. 3. Results for the *Law Students* datasets for the ranker outputs. Models marked with "n." such as ADV DR n. included a noise module in their architecture. The lines represent balanced fairness/relevance trade-offs. The dashed line represents all points with equal trade-off as the best performing comparison model, while the dotted line also takes into consideration the models we contribute.

toxic comment can be defined as having "rude, disrespectful or unreasonable" content. This dataset has been employed in fairness when evaluating both classification [14] and ranking methods [31]. The term "gay" is commonly used as a sensitive attribute, since 55% of the comments labeled toxic contain the term "gay", while only 9% of the comments which do not have the term "gay" are labeled toxic [14]. The task of interest is therefore to provide a list of comments which are ranked to most to least toxic while taking into consideration the original, biased sorting. In related work, this task has been undertaken by means of a convolutional neural network [31]. Our competing model first preprocesses the comments to generate a word representation by using a pre-trained model [30]. The obtained representation is then fed into our models. Because of the long training times for this model, we only focused on a smaller set of hyperparameters on this dataset.

C. Grid Search

We performed a simple grid search to find the best hyperparameters and architecture for our models. As is often observed in the fairness literature, removing information about the sensitive attribute often leads to decreased accuracy or relevance in the model. The same phenomenon has been observed

on representations extracted from fair neural models [28], [9]. To the end of obtaining models and representations which are both fair and relevant, we employed a metric which combines fairness and relevance for both the models and the representations extracted from them. Relevance and fairness were weighted equally. We split all the datasets with a 60/20/20 train/validation/test ratio. We then selected the models having the highest value for our metric on the validation test, and in the following we report their performance on the test set. The hyperparameter set we optimized is reported in Section 5.1. Furthermore, we tested models that had a noise module as the first layer of their architecture. For our implementation of the models and the experimental setup, see *https://zenodo.org/record/3889006*.

V. Experiments Results

In this section we present the experimental results. First, we look at the results for the different rankers on the datasets described in Section V-A. Furthermore, we compare the results of the external rankers and classifiers trained on the representations extracted by the same models in Section V-B. On the *Wiki* dataset, we also report the AUC and the GPA metrics to enable a comparison to related work that phrases fairness definitions as constrained optimization problems [31] (*Con. Opti.* in the figures.).

The comparison methods on all other datasets include: a fair listwise ranking model [44] (*DELTR*) that we also augmented with a noise module [9] (*DELTR n.*), a debiasing neural classifier [20], [9] (*Clas* and *Clas n.* if including a noise module), and an "unfair" baseline (*Base.*), which is the base DirectRanker model with $\gamma = 0$. We also employed a model that combines both the fairness mechanisms introduced in this work (*ADV FFDR, ADV FFDR n.*). These models use both the loss functions introduced in Sections III-C and III-D.

A. Model results

In the following we will discuss our results as far as the relevance (nDCG or AUC) and fairness (rND, Section IV-A1, or GPA, Section IV-A2) of the models themselves are concerned. These results can be found in Figs. 3 through 5. Note that for ease of reading, the rND and GPA metric is reported as 1-rND and 1-GPA, respectively. In all the figures. we plot the optimal line representing the model which finds the best trade-off, i.e. the smallest value of $\|(1,1) - (1-m,n)\|_1$ with $(m,n) \in \{$(rND,nDCG),(GPA,AUC)$\}$. The line for the comparison models is drawn dashed, while the one for all models (including ours) is dotted. For all datasets, we find that members of the proposed method family push the trade-off closer to the best possible model on both of the considered fairness metrics.

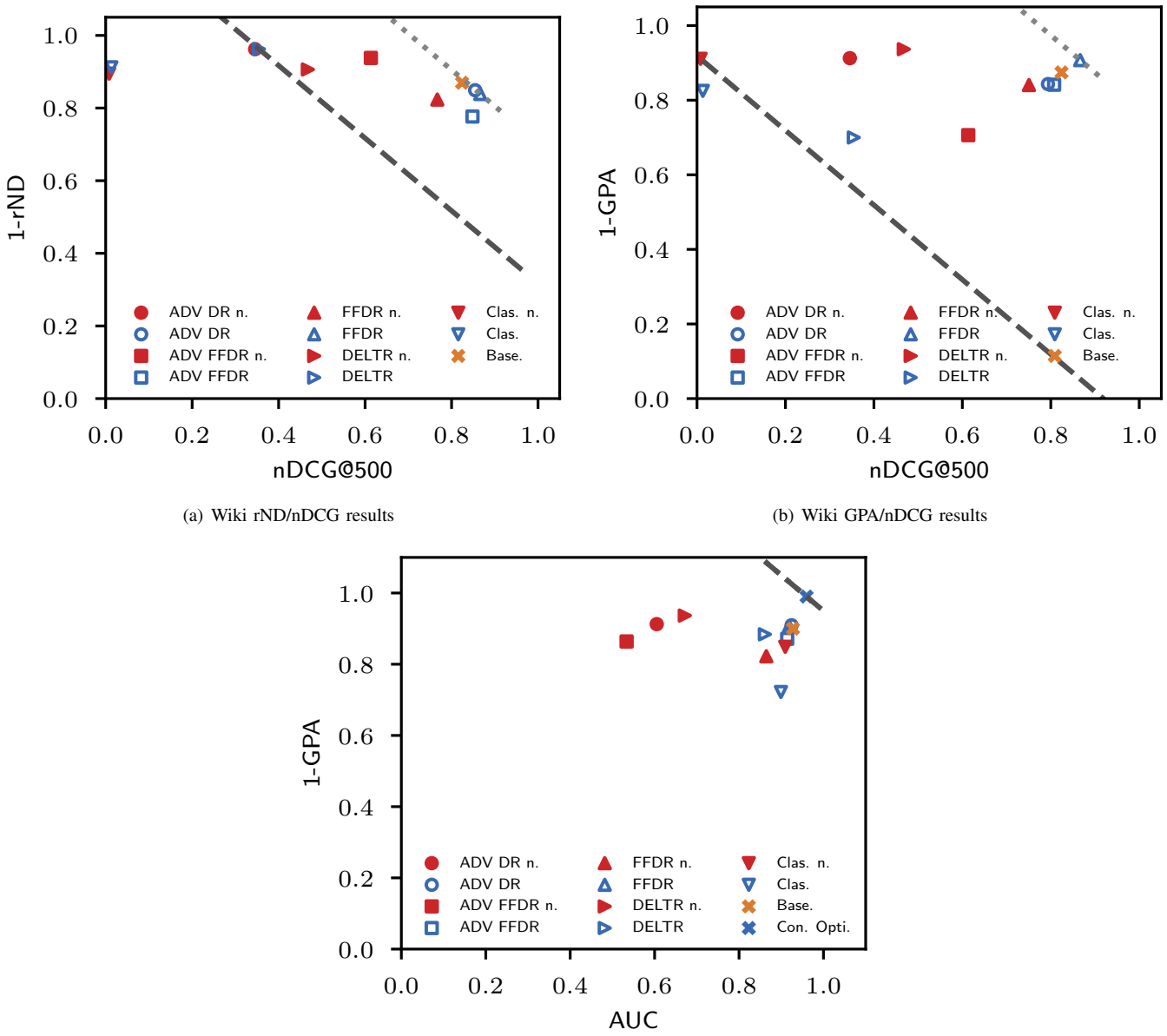
While no single model is able to outperform the comparisons on both nDCG and rND/GPA, our models are able to find strong trade-offs between relevance and fairness in all employed datasets. Overall, we found that models including the Adversarial fairness mechanism (ADV DR, ADV FFDR) performed the best. On Adult (Figs. 2(a) and 2(b)), the ADV DR model is just as fair as DELTR, while also producing rankings which are slightly more relevant. On COMPAS (Figs. 2(c) and 2(d)), our ADV DR model produces rankings which are competitive in fairness with respect to DELTR n. and the Debias Classifier. However, the output list has higher nDCG. The law dataset with gender as a sensitive attribute (Figs. 3(a) and 3(b)) provides similar insights, where, however, our best performing model employed both the mechanisms from this paper (ADV FF DR). Once again, the Debias Classifier found rankings which are fair but at the cost of slightly lowered relevance, possibly as an effect of not employing a specialized ranking loss. When considering race as the sensitive attribute (Figs. 3(c) and 3(d)), the ADV DR n. is able to retrieve the fairest rankings, while still being competitive on nDCG.

When considering the AUC/GPA trade-off on the *Wiki* dataset (Fig. 4(c)), ADV DR and FFDR have an identical performance, which is, however, surpassed by the constrained optimization model. As for the nDCG/rND trade-off (Fig. 4(a)), the ADV DR n. is once again able to find rankings which compare favorably to the others in terms of fairness, however, also having severely reduced nDCG. On this dataset, the ADV FFDR n. struck an impressive trade-off, which is about as fair as the fairest methods on the dataset (DELTR, Clas.), while also having sensibly higher nDCG. Similar observations can be made in the case of the nDCG/GPA tradeoff (Fig. 4(b)). In general, it is worthwhile to note how the DELTR model obtained sensibly fairer rankings on two of the datasets (COMPAS, law-race) when also employing a noise module, which to the best of our knowledge has not been shown before.

B. Representation results

We report results for external rankers and classifiers which have been trained on representations extracted from the employed models in Table II. We trained both linear (logistic regression) and non-linear (random forest) models and report three different metrics. To evaluate the invariance of the representation with respect to the sensitive attribute, we report classifier accuracy as the absolute distance from random guess (the majority class ratio in the dataset), which is shown as ADRG in Tabe II. Similar analyses are common in fair classification [9], [38], [28]. We transfer this analysis to the fair ranking paradigm by training rankers on the aforementioned representations. Therefore, we report the nDCG and 1-rND values for these rankers.

Experiments on the extracted representations confirm the insights derived from the model analysis, where models employing the Adversarial mechanism introduced in Section III-C were the best performing. On Adult, ADV DR and ADV FFDR are once again strong performers in all metrics, invariance included. ADV DR n. is a strong performer on COMPAS, where however DELTR n. is able to achieve better nDCG results. On law-gender ADV DR n. and ADV FF DR obtain impressive results on nDCG and rND, respectively, while still having solid performance. When instead using race as the sensitive attribute, ADV FF DR n. learns representations which have the lowest rND value, while FF DR n. displays the best nDCG with a very



(a) Wiki rND/nDCG results

(b) Wiki GPA/nDCG results

(c) Wiki AUC/GPA results

Fig. 4. Results for the *Wiki* datasets for the rankers. Models marked with n. included a noise conditioning layer. In Fig. 4(c), the values for Con. Opti. are taken from the original publication [31]. The lines represent balanced fairness/relevance trade-offs. The dashed line represents all points with equal trade-off as the best performing comparison model, while the dotted line also takes into consideration the models we contribute.

minor loss in rND. On *Wiki*, the DELTR n. model performs the best on the rND metric, however, computing outputs which are hard to sort in a relevant way. The ADV DR n. model, in comparison, has the highest value for nDCG and competitive values for rND.

VI. Discussion and Conclusions

We introduced a new family of neural network based models for the purpose of ranking *fairly* by combining the *pairwise ranking model* DirectRanker with the Gradient Reversal Layer. Combining two fairness mechanisms achieved strong results on several of the employed datasets. Employing the Adversarial mechanism described in Section III-C gave the overall best results, often when used in conjunction with the Fair Flipped mechanism. We also employed a noise module in conjunction with our own models and DELTR. Overall, members of the proposed family of fair ranking methods outperformed existing methods on all of the tested datasets. This happened consistently for two different fairness measures, although they were not directly optimized by the method (only used in model selection). Finally, we analyzed for the first time the *extracted representations* for fair ranking methods by training rankers that employ them as training vectors. When compared to other

TABLE II

PERFORMANCE OF THE REPRESENTATIONS RANKERS FOR ALL MODELS ON DIFFERENT FAIR DATASETS. THE METRICS EMPLOYED ARE, FROM LEFT TO RIGHT, ABSOLUTE DIFFERENCE TO RANDOM GUESS, 1-RND, AND THE NDCG@500. VALUES ARE MARKED BOLD IF THEY ARE THE HIGHEST ONES OF THE METRIC. MODELS MARKED WITH `n.` EMPLOYED A NOISE MODULE IN THEIR FIRST LAYER.

	COMPAS			Law-gender			Adult		
Models	ADRG	1-rND	nDCG	ADRG	1-rND	nDCG	ADRG	1-rND	nDCG
ADV DR n.	0.03	0.95	0.69	0.04	0.93	**1.00**	0.13	0.93	**0.85**
ADV DR	0.02	0.87	0.48	0.04	0.92	0.77	**0.01**	0.95	0.68
ADV FFDR n.	0.03	0.95	0.63	0.04	0.87	0.88	**0.01**	**0.98**	0.75
ADV FFDR	0.03	0.86	0.56	0.04	0.96	0.92	**0.01**	0.95	0.71
FFDR n.	0.07	0.88	0.66	0.04	0.92	0.81	**0.01**	0.95	0.77
FFDR	0.03	**0.96**	0.35	0.07	0.95	0.89	**0.01**	0.96	0.73
DELTR n.	0.03	0.95	**0.72**	0.03	0.92	0.98	0.04	0.90	0.83
DELTR	**0.01**	0.90	0.35	**0.02**	0.94	0.99	**0.01**	0.95	0.73
Clas. n.	0.04	0.86	0.45	0.04	0.91	0.97	**0.01**	**0.98**	0.56
Clas.	0.04	0.90	0.40	0.04	0.93	0.95	**0.01**	0.91	0.68
Base.	0.02	0.77	0.30	0.05	**0.97**	0.93	**0.01**	0.95	0.71

	Law-race			Wiki		
Models	ADRG	1-rND	nDCG	ADRG	1-rND	nDCG
ADV DR n.	**0.01**	0.92	0.89	0.01	0.93	**0.58**
ADV DR	0.03	0.90	0.87	**0.00**	0.94	0.51
ADV FFDR n.	0.06	**0.95**	0.95	0.01	0.93	0.17
ADV FFDR	**0.01**	0.86	0.79	**0.00**	0.92	0.04
FFDR n.	**0.01**	0.94	**1.00**	**0.00**	0.94	0.06
FFDR	0.06	0.90	0.98	**0.00**	0.90	0.08
DELTR n.	0.02	**0.95**	0.81	**0.00**	**0.95**	0.12
DELTR	0.03	0.82	0.89	**0.00**	0.89	0.55
Clas. n.	0.03	0.91	0.72	0.40	0.93	0.50
Clas.	0.20	0.93	0.68	0.02	0.91	0.55
Base.	**0.01**	0.90	0.94	**0.00**	0.92	0.22

methodologies, we have shown that our methods are able to find better trade-offs between relevance and fairness, in both disparate impact and disparate mistreatment settings.

REFERENCES

[1] Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access **6**, 52138–52160 (2018)

[2] Angwin, J., Larson, J., S.M., L, K.: Machine bias (2016)

[3] Ben-David, S., Blitzer, J., Crammer, K., Pereira, F.: Analysis of representations for domain adaptation. In: NIPS (2007)

[4] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.: Learning to rank using gradient descent. In: ICML (2005)

[5] Cao, Y., Xu, J., Liu, T.Y., Li, H., Huang, Y., Hon, H.W.: Adapting ranking svm to document retrieval. In: ACM SIGIR (2006)

[6] Cao, Z., Qin, T., Liu, T.Y., Tsai, M.F., Li, H.: Learning to rank: From pairwise approach to listwise approach. In: ICML (2007)

[7] Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: A survey on methods and metrics. Electronics **8**(8), 832 (2019)

[8] Celis, L.E., Straszak, D., Vishnoi, N.K.: Ranking with fairness constraints. In: ICALP (2018)

[9] Cerrato, M., Esposito, R., Li Puma, L.: Constraining deep representations with a noise module for fair classification. In: ACM (2020)

[10] Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data **5**(2), 153–163 (2017)

[11] Cooper, W.S., Gey, F.C., Dabney, D.P.: Probabilistic retrieval based on staged logistic regression. In: ACM SIGIR (1992)

[12] Council of European Union: Council regulation. In: (EU) no 679/2016 (2016)

[13] Dieterich, W., Mendoza, C., Brennan, T.: Compas risk scales: Demonstrating accuracy equity and predictive parity (2016)

[14] Dixon, L., Li, J., Sorensen, J., Thain, N., Vasserman, L.: Measuring and mitigating unintended bias in text classification (2018)

[15] Dua, D., Graff, C.: UCI machine learning repository (2017)

[16] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: ITCS (2012)

[17] Friedman, B., Nissenbaum, H.: Bias in computer systems. ACM TOIS **14**(3), 330–347 (1996)

[18] Friedman, J.H.: Greedy function approximation: A gradient boosting machine. Annals of Statistics **29**, 1189–1232 (2000)

[19] Fuhr, N.: Optimum polynomial retrieval functions based on the probability ranking principle. ACM TOIS **7**(3), 183–204 (1989)

[20] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**(1), 2096–2030 (2016)

[21] Kamiran, F., Calders, T.: Classifying without discriminating. 2009 2nd International Conference on Computer, Control and Communication pp. 1–6 (2009)

[22] Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems **33**(1), 1–33 (2012)

[23] Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: ECML PKDD (2012)

[24] Kohavi, R.: Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In: KDD (1996)

[25] Köppel, M., Segner, A., Wagener, M., Pensel, L., Karwath, A., Kramer, S.: Pairwise learning to rank by neural networks revisited: Reconstruction, theoretical analysis and practical performance. In: Machine Learning and Knowledge Discovery in Databases. pp. 237–252 (2020)

[26] Li, P., Wu, Q., Burges, C.J.: Mcrank: Learning to rank using multiple classification and gradient boosting. In: NIPS (2008)

[27] Liu, T.Y.: Learning to rank for information retrieval. Found. Trends Inf. Retr. **3**(3), 225–331 (2009)

[28] Louizos, C., Swersky, K., Li, Y., Welling, M., Zemel, R.: The variational fair autoencoder. preprint arXiv:1511.00830 (2015)

[29] McNamara, D., Ong, C.S., Williamson, R.C.: Provably fair representations. preprint arXiv:1710.04394 (2017)

[30] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. preprint arXiv:1301.3781 (2013)

[31] Narasimhan, H., Cotter, A., Gupta, M., Wang, S.: Pairwise fairness for ranking and regression. In: AAAI (2020)

[32] Rigutini, L., Papini, T., Maggini, M., Bianchini, M.: A neural network approach for learning object ranking. In: International Conference on Artificial Neural Networks. pp. 899–908. Springer (2008)

[33] Singh, A., Joachims, T.: Fairness of exposure in rankings. In: ACM SIGKDD 2018. Association for Computing Machinery (2018)

[34] Washington, A.L.: How to argue with an algorithm: Lessons from the compas-propublica debate. Colo. Tech. LJ **17**, 131 (2018)

[35] Western, B., Pettit, B.: Incarceration & social inequality. Daedalus **139**(3), 8–19 (2010)

[36] Wightman, L., Ramsey, H., Council, L.S.A.: LSAC national longitudinal bar passage study. LSAC research report series, Law School Admission Council (1998)

[37] Wu, Q., Burges, C.J., Svore, K.M., Gao, J.: Adapting boosting for information retrieval measures. Information Retrieval **13**, 254–270 (2010)

[38] Xie, Q., Dai, Z., Du, Y., Hovy, E., Neubig, G.: Controllable invariance through adversarial feature learning. In: NIPS (2017)

[39] Xu, J., Li, H.: Adarank: A boosting algorithm for information retrieval. In: ACM SIGIR (2007)

[40] Yang, K., Stoyanovich, J.: Measuring fairness in ranked outputs. In: SSDBM '17. Association for Computing Machinery, New York, NY, USA (2017)

[41] Yang, K., Stoyanovich, J., Asudeh, A., Howe, B., Jagadish, H., Miklau, G.: A nutritional label for rankings. In: Proceedings of the 2018 international conference on management of data. pp. 1773–1776 (2018)

[42] Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: WWW (2017)

[43] Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., Baeza-Yates, R.: Fa*ir. In: CIKM (2017)

[44] Zehlike, M., Castillo, C.: Reducing disparate exposure in ranking: A learning to rank approach. preprint arXiv:1805.08716 (2018)

[45] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: ICML (2013)