Via VIII Febbraio 1848, 2 - 35122, Padova
CF 92315730280 | **Cap.Soc**. Euro 100.000,00 i.v.
Email: info@consorzioinest.it
PEC: consorzio_inest@pec.it

Spoke 6

—

TOURISM, CULTURE AND CREATIVE INDUSTRIES

Leader: Ca' Foscari University of Venice

—

Research Topics

RT1
New Digital Technologies
RT2
Data Analytics
RT3
Sustainable Business Models
RT4
New Narratives and Communication Strategies

iNEST Interconnected Nord-Est Innovation Ecosystem

Via VIII Febbraio 1848, 2 - 35122, Padova
CF 92315730280 | **Cap.Soc**. Euro 100.000,00 i.v.
Email: info@consorzioinest.it
PEC: consorzio_inest@pec.it

Task S6_RT1.3 – Deliverable DS6_RT1.3.4 - Report on novel methodologies for optimal recommendation by empowering Machine Learning algorithms and Learning-to-Rank models

1 INTRODUCTION

This document presents the research and results developed during my research grant (*Assegno di Ricerca*) for the iNEST project, specifically within the scope of DS6_RT1.3.4, titled *Report on novel methodologies for optimal recommendation by empowering Machine Learning algorithms and Learning-to-Rank models*.

The core of the research focused on the design and development of vertical search engines dedicated to tourist destinations, accommodation facilities, and territorial information. These systems were required to incorporate aspects of fairness and/or explainability.

In the context of search engines, and more broadly Information Retrieval (IR), fairness refers to the balanced treatment of items in rankings, avoiding discrimination against entities belonging to sensitive or minority groups. Fairness is closely related to the datasets used to train machine learning (ML) models, which may contain sampling biases or reflect social prejudices. This can lead models to generate rankings that unfairly disadvantage certain user groups (e.g., based on gender) or categories of tourist services (e.g., hotels versus apartments).

Explainability in IR refers to the ability to explain how decisions are made by the ML-based ranking models. The aim is to provide a simple and intuitive link between the input features and the model's relevance predictions, highlighting which aspects of the query, user profile, or returned item influenced the ranking.

In summary, fairness helps build trust among stakeholders, such as hospitality providers who want assurance that their offerings are ranked fairly. Explainability, on the other hand, can improve the experience for tourists and help businesses understand which features they may need to improve in order to gain more visibility.

During the research grant, I extensively analyzed fairness in Learning-to-Rank (LTR), which led to the publication of two project milestones: **LambdaRank Gradients are Incoherent** [16, 19] and **LambdaFair: a Fair and Effective LambdaMART** [20, 21].

2 STATE OF THE ART

Current search engine technologies rely on machine learning algorithms, particularly LTR models [3, 4, 28], to re-rank results presented to users. These systems leverage complex features related to the query and user context, as well as characteristics of the items being ranked (e.g., documents, points of interest), to identify the most relevant results that fulfill users' information needs.

iNEST Interconnected Nord-Est Innovation Ecosystem

Via VIII Febbraio 1848, 2 - 35122, Padova
CF 92315730280 | **Cap.Soc**. Euro 100.000,00 i.v.
Email: info@consorzioinest.it
PEC: consorzio_inest@pec.it

In terms of fairness in LTR models, the literature primarily categorizes existing approaches into three broad groups: *pre-processing*, *in-processing*, and *post-processing* methods [34,35].

Pre-processing methods aim to mitigate bias in the training data prior to model learning. These techniques are generally model-agnostic and focus on enhancing individual fairness without modifying the underlying ML algorithms. For instance, [15] proposes transforming feature vectors into more equitable representations to improve individual fairness. Similarly, [36] introduces a strategy that balances the dual objectives of maintaining an accurate representation of the data and hiding sensitive group membership information to enhance fairness.

In-processing methods incorporate fairness constraints directly into the learning process, typically by modifying the loss function or adding regularization terms that encourage fairness. A notable example is DELTR [32], which reduces group-level disparities in item exposure within rankings while maintaining relevance. Similarly, Fair-PG-Rank [26] addresses fairness by modeling exposure as expected attention, operating under a merit-based constraint that ensures items receive exposure proportional to their relevance, thus balancing fairness and effectiveness. Recent research has identified stochastic Plackett-Luce (PL) ranking models [18,24] as a powerful in-processing approach for jointly optimizing effectiveness and fairness metrics. Unlike deterministic algorithms that depend on heuristic optimization, PL models are fully differentiable and can be optimized via stochastic gradient descent directly on ranking metrics. However, estimating gradients in practice is challenging because it involves summing over all possible permutations of items. To tackle this, Oosterhuis [22] proposed PL-Rank, an efficient method to estimate gradients of PL models with respect to both fairness and effectiveness metrics by leveraging specific properties of ranking metrics and PL distributions. Further advancing this line of work, the PL-Rank-3 algorithm [23] introduced by Oosterhuis achieves unbiased gradient estimates with computational efficiency comparable to state-of-the-art sorting algorithms. Building upon these advances, Gorantla *et al.* [11] proposed Group-Fair-PL, which optimizes group fairness through a novel objective that computes expected ranking utility over only those rankings satisfying explicit group representation constraints. Group-Fair-PL enforces either equal or proportional representation of protected items within the top-k ranks. Equal representation requires the same number of items from each group in the top-k, whereas proportional representation reflects the relative group proportions in the dataset.

Post-processing methods modify the output rankings after model inference to satisfy fairness criteria. FA*IR [31] adjusts the ranking to ensure that the proportion of protected candidates meets minimum thresholds at different ranking depths. The CFAΘ algorithm [33] provides a mechanism to continuously trade off between individual and group fairness, enabling practitioners to fine-tune the fairness criteria applied to the final ranking output.

iNEST Interconnected Nord-Est Innovation Ecosystem

Via VIII Febbraio 1848, 2 - 35122, Padova
CF 92315730280 | **Cap.Soc.** Euro 100.000,00 i.v.
Email: info@consorzioinest.it
PEC: consorzio_inest@pec.it

3 BACKGROUND

Learning-to-rank methods are a cornerstone of information retrieval systems, used to re-order a set of documents in response to a query. Given a query q and a candidate document set $D = \{d_1, \dots, d_n\}$, an LTR model learns a scoring function that assigns each document $d_i \in D$ a score s_i. The final ranking π is obtained by sorting documents in descending order of s_i. Formally, π is a permutation of $\{1, \dots, n\}$ such that $\pi[r] = i$ means d_i occupies the r-th position, where $r \in \mathbb{N}^+$ and $1 \leq r \leq n$. For any $r_\ell \leq r_u$, we write

$$\pi[r_\ell, r_u]$$

to denote the subsequence of documents ranked from position r_ℓ to r_u. In particular, the prefix of length r_u is $\pi[1, r_u]$. Finally, we use

$$d_i \prec_\pi d_j$$

to indicate that d_i appears before d_j in the ranking π.

3.1 EFFECTIVENESS METRIC

We evaluate both effectiveness and fairness. For effectiveness, we use the *Normalized Discounted Cumulative Gain* (NDCG) metric [13]. Let π be a ranking of D of size n, and let each document $d_i \in D$ have a relevance label y_i. The Discounted Cumulative Gain part of NDCG is defined as:

$$\mathrm{DCG}(\pi) = \sum_{r=1}^{n} \frac{2^{y_{\pi[r]}} - 1}{\log_2(r+1)} \quad \text{and} \quad \mathrm{IDCG}(\pi) = \max_{\pi'} \mathrm{DCG}(\pi').$$

Then, the NDCG metric is defined as:

$$\mathrm{NDCG}(\pi) = \frac{\mathrm{DCG}(\pi)}{\mathrm{IDCG}(\pi)},$$

which yields a value in $[0, 1]$, with higher values indicating better alignment with the true relevance ordering. To reflect typical user behavior, where only the top results are examined, NDCG is computed at cutoff k, denoted NDCG@k.

3.2 FAIRNESS METRIC

In [20, 21], the fairness metric we employed is *Normalized Discounted Difference* (rND) [30], which captures group fairness as statistical parity [34]. Given a ranking π of D with $|\mathcal{G}^+|$ protected items, rND evaluates whether, for various cutoffs r, the top-r positions contain a fraction of protected items equal to $|\mathcal{G}^+|/n$. Here $\mathcal{G}^+ \subseteq D$ denotes the protected subset.

iNEST Interconnected Nord-Est Innovation Ecosystem

Via VIII Febbraio 1848, 2 - 35122, Padova
CF 92315730280 | **Cap.Soc.** Euro 100.000,00 i.v.
Email: info@consorzioinest.it
PEC: consorzio_inest@pec.it

rND divides the ranking into overlapping prefixes of lengths $r = b, 2b, 3b, \dots$, where $b > 1$ is the *bin size*. For each prefix length r, it computes the absolute deviation from the ideal proportion and discounts it by $\log_2(r)$. The rND metric is defined as follows:

$$\text{rND}(\pi) = \frac{1}{\text{rDmax}} \sum_{r=b,2b,\dots}^{n} \frac{1}{\log_2(r)} \left| \frac{|\mathcal{G}^+_{\pi[1,r]}|}{r} - \frac{|\mathcal{G}^+_{\pi[1,n]}|}{n} \right| ,$$

where $\mathcal{G}^+_{\pi[1,r]}$ is the set of protected items in the top-r, and rDmax is the worst-case sum obtained by ranking all elements of the group with smaller cardinality first. The metric lies in $[0, 1]$, with lower values indicating greater fairness. As with NDCG, the rND metric can be computed up to a certain cutoff k, i.e., rND@k.

3.3 DATASETS

We performed an extensive evaluation on publicly available datasets. In **LambdaRank Gradients are Incoherent** [16,19], we used the following three datasets. The ISTELLA-X [17] dataset has the highest average number of documents per query and the largest number of non-relevant documents. YAHOO! LEARNING TO RANK CHALLENGE SET 1 [6] dataset is the smallest dataset, with about $700,000$ documents and an average of 23.73 documents per query. The MSLR Web30K Fold 1 [25] dataset consists of feature vectors extracted from query-document pairs. It is the most balanced dataset, with about half of the documents labeled as relevant. All datasets have graded 5-level relevance labels ranging from 0 (non-relevant) to 4 (highly relevant).

For **LambdaFair: a Fair and Effective LambdaMART** [20, 21] instead, we considered three additional datasets widely used in fairness-aware learning-to-rank research: Statlog (German Credit Data) [12], and Home Mortgage Disclosure Act (Connecticut) [8]. The German Credit Data dataset contains binary relevance labels related to creditworthiness, with $1,000$ individuals per query sampled to create $100,000$ queries. We derived two variants of this dataset by splitting protected and unprotected groups based on:

- **Sex**: females as protected, males as unprotected [14, 26, 29],
- **Age**: individuals under 35 as protected, those 35 and older as unprotected [1].

The Home Mortgage Disclosure Act (Connecticut) dataset includes home mortgage loan records across US states since 2007. We focused on Connecticut, using data from 2013–2015 for training, 2016 for validation, and 2017 for testing [11]. Similar to German Credit Data, we sampled 50 individuals per query with a $4{:}1$ ratio of non-approved to approved loans, for a total of $100,000$ queries. The protected group consists of females, and the unprotected group of males.

iNEST Interconnected Nord-Est Innovation Ecosystem

Via VIII Febbraio 1848, 2 - 35122, Padova
CF 92315730280 | **Cap.Soc.** Euro 100.000,00 i.v.
Email: info@consorzioinest.it
PEC: consorzio_inest@pec.it

Since MSLR-30K does not provide protected group labels, we followed previous works [1, 14, 27, 29] and used the QualityScore2 feature (QS2, feature ID 133) as a discriminatory attribute. Following Vardasbi *et al.* [27], documents with QS2 values below 10 are assigned to the protected group, and those with QS2 values equal or above 10 to the unprotected group.

All datasets are partitioned into training, validation, and test sets following a 60%-20%-20% split.

3.4 LAMBDAMART

LambdaMART [3] is a learning algorithm widely used in IR to train effective ranking models by directly optimizing a target metric. It addresses the non-differentiability and flat regions of ranking metrics [2, 4, 9, 19] by leveraging a smooth gradient approximation.

Formally, let q be a query, $D = \{d_1, \dots, d_n\}$ the candidate documents, and $Y = \{y_1, \dots, y_n\}$ their relevance labels. LambdaMART constructs a ground-truth pairwise preference set P where $(i, j) \in P$ if and only if $y_i > y_j$. For each document d_i, LambdaMART computes an approximated gradient λ_i^Z for the ranking metric Z (e.g., NDCG@k) as:

$$\lambda_i^Z = \sum_{j\,:\,(i,j)\in P} \lambda_{ij}^Z - \sum_{k\,:\,(k,i)\in P} \lambda_{ki}^Z, \tag{1}$$

where each partial approximated gradient λ_{ij}^Z is defined as:

$$\lambda_{ij}^Z = \frac{\partial C(s_i - s_j)}{\partial s_i} = \frac{-\sigma}{1 + e^{\sigma(s_i - s_j)}} |\Delta Z_{ij}|. \tag{2}$$

Here, s_i and s_j are the scores predicted for d_i and d_j, C is the RankNet cost function [5], and

$$\Delta Z_{ij} = Z(\pi_{ji}) - Z(\pi_{ij})$$

denotes the change in metric Z when swapping d_i and d_j in the current ranking. The scalar σ controls the sigmoid's steepness.

These λ_i^Z values serve as pseudo-gradients in iterative optimization algorithms such as neural networks (e.g., LambdaRank [4]) or gradient-boosted trees (e.g., LambdaMART [3]).

4 CONTRIBUTION

The primary contribution of the research conducted during the grant period was centered on the development of LTR algorithms with an emphasis on fairness, particularly through in-processing methods. This line of work addresses a critical issue

iNEST Interconnected Nord-Est Innovation Ecosystem

Via VIII Febbraio 1848, 2 - 35122, Padova
CF 92315730280 | **Cap.Soc**. Euro 100.000,00 i.v.
Email: info@consorzioinest.it
PEC: consorzio_inest@pec.it

in the deployment of modern search and recommendation systems: the risk of amplifying biases and producing systematically unfair rankings.

The research aimed to integrate fairness considerations directly into the optimization process of LTR models, rather than relying on pre-processing (data transformation) or post-processing (re-ranking) techniques. This in-processing approach is both theoretically and practically significant, as it allows for fairness constraints or objectives to be embedded within the model's learning dynamics, potentially leading to more balanced and equitable outputs without sacrificing effectiveness. Two main contributions were achieved during the grant period, each resulting in peer-reviewed publications:

- **LambdaRank Gradients are Incoherent** [16, 19]: This work revisits the foundational LambdaRank framework, uncovering a mathematical incoherence in the gradient formulation used for training that can unfairly misrank equally relevant items. By analyzing the gradient dynamics, the research demonstrates that the implicit assumptions made by LambdaRank can lead to unintended optimization behaviors. The paper proposes a refined understanding of ranking gradients, paving the way for more principled and fair LTR methods.
- **LambdaFair: a Fair and Effective LambdaMART** [20, 21]: Building on the insights above, this paper introduces LambdaFair, an extension of LambdaMART that incorporates fairness constraints directly into the learning process. LambdaFair is designed to mitigate exposure disparity among groups while maintaining competitive ranking performance. Extensive experimental evaluation on publicly available datasets confirms that LambdaFair achieves a better trade-off between fairness and effectiveness compared to existing baselines.

Overall, the research contributes both theoretical insights and practical algorithms that advance the state of the art in fair information retrieval. It demonstrates that it is possible to reconcile the goals of effectiveness and fairness in ranking, and it provides tools that can be adopted or extended in real-world systems.

4.1 LAMBDARANK GRADIENTS ARE INCOHERENT

Many LTR algorithms still rely on gradient-based optimization, either by approximating the ranking metric or by constructing heuristic gradients, as these methods have proven to be highly effective in practice. As mentioned above, LambdaMART optimizes a non-differentiable objective by generating ad hoc gradients for each document. These gradients are based on heuristic assumptions about each document's contribution to the overall ranking and its interactions with other documents. Consequently, the gradients produced by LambdaMART are inherently approximate.

iNEST Interconnected Nord-Est Innovation Ecosystem
Via VIII Febbraio 1848, 2 - 35122, Padova
CF 92315730280 | **Cap.Soc**. Euro 100.000,00 i.v.
Email: info@consorzioinest.it
PEC: consorzio_inest@pec.it

In [16, 19], we demonstrate that LambdaMART and its related methods, such as LambdaRank [4] and the loss functions introduced by [28], exhibit an intrinsic issue stemming from their heuristic foundations: the generation of incoherent gradients. Our analysis revealed three critical and previously undocumented behaviors in LambdaMART and its derivatives:

- *i)* We found that LambdaMART suffers from gradient incoherencies that hinder the learning process. Specifically, it can assign stronger downward gradient forces to documents with higher relevance than to those with lower relevance. This leads the model to mislearn the correct ranking order.
- *ii)* We observed that optimizing truncated IR metrics exacerbates these incoherencies, further degrading model performance. Although truncated metrics (e.g., top-k) are useful for focusing learning on the top ranks and reducing training time, they also increase the risk of incorrect gradient estimation.
- *iii)* Finally, we discovered that optimizing truncated metrics introduces unfair treatment of equally relevant documents. In particular, when two or more documents have the same relevance, those ranked lower in the list receive weaker upward gradient signals than those in higher positions. This puts lower-ranked yet equally relevant documents at a disadvantage, since they require a stronger push to improve their rank.

These findings reveal that the widely used LambdaMART algorithm and its derivatives can exhibit unfair behavior, especially in scenarios where fairness across equally relevant items is essential.

4.1.1 GRADIENT INCOHERENCY AND UNFAIR DOCUMENT COMPARISON

Gradient-based learning algorithms, such as artificial neural networks or gradient-boosted decision trees, run iterative updates to build a ranker that minimizes a given cost function C. For instance, gradient-boosted decision trees iteratively learn a new tree that approximates $\partial C / \partial s_i$ for each document d_i in the training set D and its score s_i. Unfortunately, most IR metrics are rank-based: they depend on ranking π rather than on s_i. This makes the cost function either flat or non-differentiable. Note that π is the ranking over the documents $d_i \in D$ sorted in decreasing order of scores s_i predicted by the ranker, and $\pi[i]$ denotes the position of document d_i in the ranking.

LambdaRank's cost function defined by [4] is one of the most relevant approaches used to tackle this problem, and it stems from the RankNet cost proposed by [5], which is enhanced by considering the impact on the IR metric. The gradient is computed on the basis of pair-wise lambdas λ_{ij} as defined in Equation 1, where P is the set of ordered documents pairs (i, j) such that $y_i > y_j$, i.e., $P = \{(i, j) \mid d_i, d_j \in D \ \wedge \ y_i > y_j\}$. The value of λ_{ij} estimates the change on the

iNEST Interconnected Nord-Est Innovation Ecosystem

Via VIII Febbraio 1848, 2 - 35122, Padova
CF 92315730280 | **Cap.Soc.** Euro 100.000,00 i.v.
Email: info@consorzioinest.it
PEC: consorzio_inest@pec.it

Table 1: Detailed computation of LambdaMART gradients.

d_i	$\pi[i]$	y_i	s_i	λ_i	
d_1	1	4	0.02	$\lambda_1 = \lambda_{12} + \lambda_{13}$	$\approx 0.176 + 0.221 \approx 0.397$
d_2	2	0	0.01	$\lambda_2 = -\lambda_{12} - \lambda_{32}$	$\approx -0.176 - 0.004 \approx -0.180$
d_3	3	1	0.00	$\lambda_3 = -\lambda_{13} + \lambda_{32}$	$\approx -0.221 + 0.004 \approx -0.217$

cost function C when the difference between the two scores s_i and s_j increases or decreases.

To manage user behavior and increase training efficiency, real-world applications of information retrieval systems mostly try to optimize the effectiveness only for the first k results. IR metrics naturally provide a truncated version, i.e. NDCG@k is computed by considering only the contribution of the top-k ranked documents.

By training the model to optimize a truncated metric Z to a certain truncation level τ, pairs of documents ranked beyond τ are not considered since the corresponding contribution to the metric is equal to 0. Thus, in order to reduce the training time, the number of document pairs in P is limited while computing the gradients λ_i in Equation 1 by replacing the set P with $I_\tau = \{(i,j)|d_i, d_j \in D \wedge y_i > y_j \wedge \min(\pi[i], \pi[j]) \leq \tau\}$.

It is important to note that, although closely related, the truncation level τ is different from the metric cutoff k. The former affects the number of document pairs to process, and the latter affects the evaluation of the metric. Moreover, they may not be equal, i.e. τ may be slightly larger than k to process more pairs during the training phase.

Table 1 shows an example of LambdaMART gradients when maximizing NDCG. The query has only three documents with their ranks $\pi[i]$ and scores s_i predicted by the model, and relevance label y_i. The top-ranked document with relevance equal to 4 and is correctly pushed up by the gradient λ_1. Interestingly enough, the second and third documents are misranked with labels 0 and 1 respectively. The LambdaMART gradient is negative for both documents, but the document with the larger label is pushed down with greater strength. We may conclude that such gradients are not going to improve the ranking but rather increase the gap between the two misranked documents. We call this phenomenon *gradients incoherency*.

To explain in detail the reason for such behavior, in Table 1 we report the computation of the document gradients λ_i as a function of the pair-wise λ_{ij} according to Equation 1 in case of the NDCG metric. Document d_1 has a positive gradient λ_1 as it is ranked higher than documents with smaller relevance labels. Document d_2 is the least relevant and receives a negative gradient contribution from both the other documents. Unexpectedly, document d_3 receives the strongest downward push even if it has a higher label than d_2. The reason is that swapping document d_1 with d_3 has a larger impact on the NDCG than swapping d_1 with d_2, resulting in $\lambda_{13} > \lambda_{12}$. LambdaMART prefers avoiding the risk of moving d_1 to the third position rather than pushing d_3 up to the second place. Indeed, this comes from the discount factor of NDCG metric that demotes documents' contributions in the lower ranks.

iNEST Interconnected Nord-Est Innovation Ecosystem
Via VIII Febbraio 1848, 2 - 35122, Padova
CF 92315730280 | **Cap.Soc**. Euro 100.000,00 i.v.
Email: info@consorzioinest.it
PEC: consorzio_inest@pec.it

These gradients clearly push the ranking away from the ideal configuration as we would prefer having λ_3 larger than λ_2.

The phenomenon of gradient incoherencies is significantly further exacerbated when optimizing truncated metrics. This is because by removing (i, j) pairs from set P, the λ_i gradients of the relevant documents under τ will be incomplete since they will lose some λ_{ij}. As a result, relevant documents below τ will receive even less upward push.

Moreover, the substitution of the set P with I_τ introduces an unfair optimization process among equally relevant documents. In fact, while optimizing truncated metric, two equally relevant documents can receive a different number of λ_{ij} contributions, not experienced during un-truncated metric optimization.

In Figure 1, it is possible to see a clear example of unfair document comparison introduced by truncated metric optimization. In the example, a truncated metric optimization is forced by implying $\tau = 1$. When truncated metric optimization is used the document in position 1 with relevance 2 receives more contribution from the other documents than the document in position 2 while being equally relevant. In fact, the document in position 2 receives contribution $\lambda_i j = 0$ from all documents below the truncation level, while the document in position 1 receives contribution $\lambda_i j \geq 0$.

4.1.2 LAMBDA-*EX*

The first contribution of the research grant is Lambda-e*X*, a learning algorithm able to cancel the gradient incoherency exacerbation and avoid the introduction of unfair document comparisons. We claim that the exacerbation of the incoherencies and the unfair comparisons are due to missing computations of the λ_{ij} gradients. More specifically, relevant documents that are not ranked above the truncation level are not evaluated against all the candidate documents of the query but only against the top-k, and this ignores some of the λ_{ij} and under-estimates their gradient.

To keep the training focused on user behavior and tackle the problem of gradient incoherencies and unfair document comparison, we designed Lambda-e*X*, which *extends* the set of document pairs considered by LambdaMART when computing gradients. To achieve this goal, we define a set of documents $X \subseteq D$ so as to include in X all the documents for which we want a *complete* gradient estimation, i.e., those that deserve or not to be ranked at the top positions. Moreover, the set $\mathcal{X}$ contains the documents for which we want a fair comparison. As mentioned, X is a subset of D, so not all the documents in D are provided with a fair comparison. However, due to a limited number of positions effectively viewed by the users, i.e., the top-k positions, we do not require that all documents have a fair comparison, but just those that should be ranked in the top-k position to maximize the user need, i.e., the relevance to the user query.

Lambda-e*X* thus computes each λ_i gradient as in Equation 1 but basing on the

iNEST Interconnected Nord-Est Innovation Ecosystem

Via VIII Febbraio 1848, 2 - 35122, Padova
CF 92315730280 | **Cap.Soc.** Euro 100.000,00 i.v.
Email: info@consorzioinest.it
PEC: consorzio_inest@pec.it

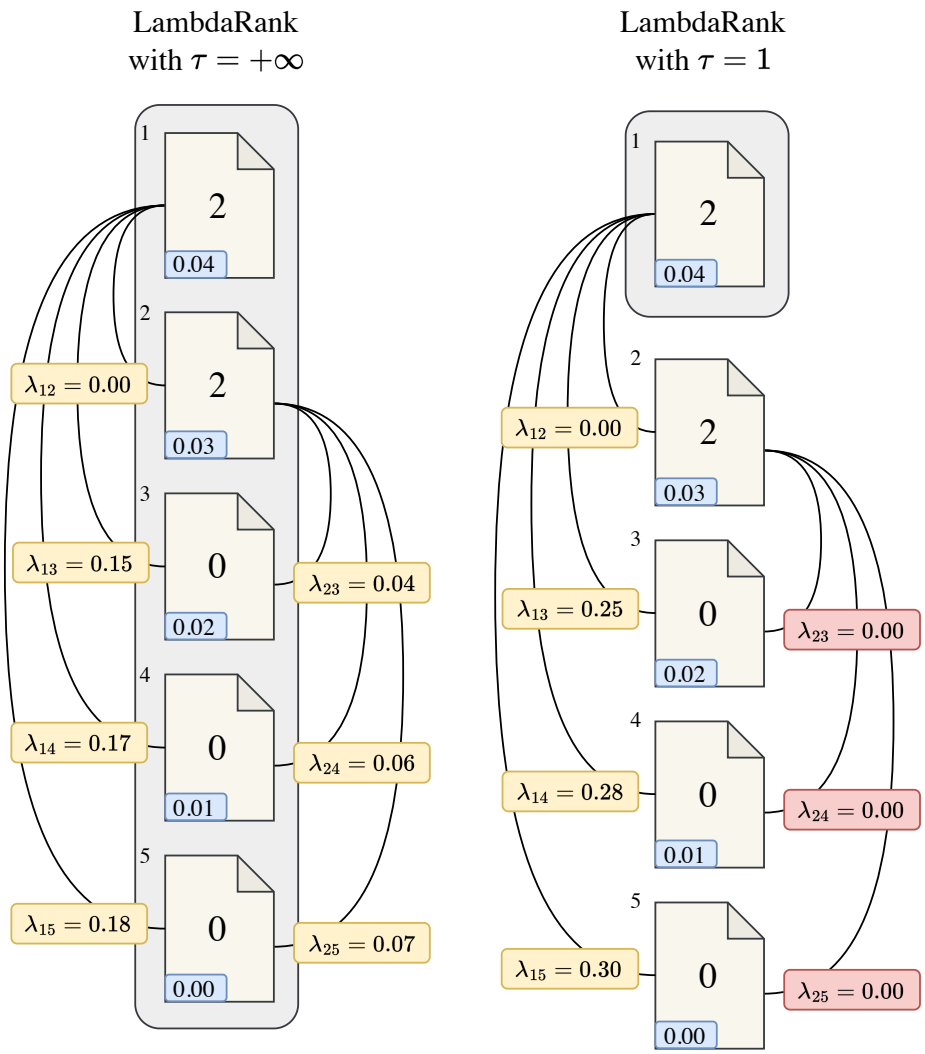

Figure 1: Examples of unfair document comparison. When truncated metric optimization is used (e.g., $\tau = 1$), the document in position 1 with relevance 2 receives more contribution from the other documents than the document in position 2 while being equally relevant.

set:

$$I_X = \{(i,j) \mid d_i, d_j \in D \wedge y_i > y_j \wedge (d_i \in X \vee d_j \in X)\}$$

How does Lambda-*eX* populate the set X? Let k be the cutoff of the IR metric being optimized. Lambda-*eX* includes in X all the documents ranked in the top-k positions by the current model and the contender documents below the cutoff k. The contender documents are those not placed in the top-k positions but required to maximize the ranking metric. Since the number of contender documents could be large, to limit the size of X to about k, Lambda-*eX* uses some criteria to select the

iNEST Interconnected Nord-Est Innovation Ecosystem

Via VIII Febbraio 1848, 2 - 35122, Padova
CF 92315730280 | **Cap.Soc**. Euro 100.000,00 i.v.
Email: info@consorzioinest.it
PEC: consorzio_inest@pec.it

contender documents to include in X. We propose three different ways to select the contender documents.

- `static`: let h be the number of documents that are required to maximize the metric Z, but the model did not rank among the top-k positions. This strategy adds in X the h most relevant documents not yet in the top-k. This strategy provides a partial fair document compassion since it compares the minimal number of documents that deserve to be in the top-k in order minimize the training time, i.e., the training efficiency.
- `random`: analogous to `static`, but ties are broken randomly instead of rank-based. A random selection allows the model to be fairer since it compares all contenders during training and improves model generalization.
- `all`: analogous to `static`, but ties are not broken. If there are more than h documents with the desired relevant labels, they are all included in X. This enhances generalization and fairness at the expense of efficiency. With the `all` strategy, all documents that should be ranked in the top-k are compared with the others, so providing the fairest strategy among the three.

We also propose two hybrid variants: `all-static` and `all-random`. The goal is to limit the size of X. Depending on the query, the `all` may potentially include all the relevant documents. To avoid such blow-up of X, the hybrid strategies roll back to either `static` or `random` in this degenerate case; otherwise, they implement the `all` strategy.

4.1.3 RESULTS

To evaluate the effectiveness of the proposed approach, we defined two baselines: LambdaMART that optimizes the truncated version of NDCG metric with truncation level τ equal to the metric cutoff k and the other with $\tau = k + 3$ as suggested in [7]. We compared the baselines with each version of Lambda-*eX* defined in Section 4.1.2. In Table 2, we summarize the performance for each model. The results show how Lambda-*eX* obtains statistical significance improvements by overcoming the exacerbation of gradient incoherencies introduced by LambdaMART when optimizing a truncated metric.

4.2 LAMBDAFAIR: A FAIR AND EFFECTIVE LAMBDAMART

In [20, 21], we introduce LambdaFair, an in-processing method based on LambdaMART, designed to jointly optimize two ranking metrics: NDCG (subsection 3.1) and rND (subsection 3.2). Leveraging the gradient formulation of

iNEST Interconnected Nord-Est Innovation Ecosystem

Via VIII Febbraio 1848, 2 - 35122, Padova
CF 92315730280 | **Cap.Soc.** Euro 100.000,00 i.v.
Email: info@consorzioinest.it
PEC: consorzio_inest@pec.it

Table 2: Statistically significant improvement w.r.t. LambdaMART$_{\tau=k+3}$ according to Fisher's randomization test defined by [10] (with a two-sided p-value) are marked with bold ($p = 0.05$) and bold-italic ($p = 0.01$).

dataset	NDCG@k k	LambdaMART $\tau=k$	LambdaMART $\tau=k+3$	Lambda-*eX* static	Lambda-*eX* random	Lambda-*eX* all	Lambda-*eX* all-static	Lambda-*eX* all-random
ISTELLA-X	5	73.32	75.35**	75.19**	75.17**	75.15**	75.19**	75.17**
ISTELLA-S	5	70.19	70.64*	70.67**	70.71**	70.55	70.65*	70.64*
ISTELLA-F	5	67.02	67.62**	67.55**	67.67**	67.50**	67.68**	67.71**
YAHOO! SET 1	5	75.35	75.85**	75.67	75.59	75.63	75.73	75.67
MSLR-30K	5	50.66	51.22	50.95	50.96	51.24	51.42*	51.38
ISTELLA-X	10	77.53	78.61	78.61	78.61	78.61	78.61	78.61
ISTELLA-S	10	76.35	76.71**	76.66**	76.70**	76.69**	76.72**	76.70**
ISTELLA-F	10	71.85	72.39**	72.42**	72.46**	72.42**	72.35**	72.46**
YAHOO! SET 1	10	79.62	79.84	79.66	79.75	79.78	79.81	79.80
MSLR-30K	10	52.66	52.98	52.96	53.08	53.23*	53.19	53.14
ISTELLA-X	15	79.00	79.45	79.44	79.48	79.44	79.44	79.48
ISTELLA-S	15	80.63	80.73*	80.69	80.71*	80.75**	80.80**	80.73*
ISTELLA-F	15	75.46	75.87**	75.94**	75.90**	75.92**	76.00**	76.00**
YAHOO! SET 1	15	82.01	82.03	81.94	82.07	82.04	82.07	82.04
MSLR-30K	15	54.60	54.67	54.82	54.93**	54.84	54.75	54.83

LambdaMART, this joint optimization is achieved via a convex combination, as follows:

$$\lambda_i = \alpha\lambda(E)_i^{\text{NDCG}} + (1-\alpha)\lambda(F)_i^{\text{rND}} \ ,$$

where E and F denote the sets of pairwise preferences associated with the NDCG (effectiveness) and rND (fairness) metrics, respectively. The hyper-parameter α controls the trade-off between the two objectives. Note that E is built from the ground-truth relevance labels and corresponds to the set P used in Equation 1.

Although the definition is straightforward, jointly optimizing these two metrics presents subtle challenges that must be addressed with care. For NDCG, the preference set E is directly derived from the relevance labels associated with each document under a query. These labels induce a *partial order* over the document set D, and thus E includes all ordered pairs (i, j) satisfying $y_i > y_j$, which suffice to optimize NDCG.

In contrast, optimizing rND requires that every ranking prefix reflects the same protected group proportion as the entire ranking, i.e., $|\mathcal{G}^+|/n$. Consequently, no inherent partial order exists on D from which to derive F.

Moreover, treating sets E and F independently can lead to sub-optimal results for both metrics. Consider the toy example in Figure 2. From the NDCG perspective, the relevant document d_5 must appear first, hence $E = \{(5, i) \mid i \neq 5\}$. From the fairness perspective, assuming an rND bin size $b = 3$, the ideal configuration requires one protected item in each bin, one in the top-3, the other in the bottom-3. To enforce

iNEST Interconnected Nord-Est Innovation Ecosystem

Via VIII Febbraio 1848, 2 - 35122, Padova
CF 92315730280 | **Cap.Soc**. Euro 100.000,00 i.v.
Email: info@consorzioinest.it
PEC: consorzio_inest@pec.it

document	d_1	d_2	d_3	d_4	d_5	d_6
rank	1	2	3	4	5	6
relevance	0	0	0	0	1	0
group	-	-	-	-	+	+

Figure 2: Toy example to discuss joint metric optimization. Six documents $d_1, \dots, d_6$, with different relevance and group membership, with $\mathcal{G}^+ = \{d_5, d_6\}$. Note that d_5 is relevant and also belongs to the protected group.

this, we could construct two distinct preference sets: $F' = \{(5, i) \mid i \neq 5 \wedge i \neq 6\}$ and $F'' = \{(6, i) \mid i \neq 6 \wedge i \neq 5\}$. F' pushes d_5 into the first bin and leaves d_6 in the second, while F'' does the reverse.

Both F' and F'' are equally valid from a fairness perspective. However, F' is aligned with E, as both promote d_5, leading to the ranking $[5, 1, 2, 3, 4, 6]$, which is optimal for both metrics. Conversely, combining E with F'' yields the ranking $[5, 6, 1, 2, 3, 4]$, where both protected items appear in the first bin, resulting in a sub-optimal rND score.

This example highlights that the definition of pairwise preferences is crucial: preference pairs for one metric cannot be designed in isolation from the other.

4.2.1 STRATEGIES

This section presents three heuristic strategies to compute the set F for the fairness-oriented ranking metric rND, using the fixed set E, defined for optimizing NDCG, as a reference. In all strategies described below, the set F is dynamically computed at each step where the learning algorithm estimates the lambdas, unlike the NDCG set E, which is precomputed once.

Before detailing how F is constructed in LambdaFair, it is important to recall that LambdaMART-based algorithms estimate the impact of swapping document positions in the ranking π to compute metric gradients during learning. We define ΔrND_{ij} as the change in rND resulting from swapping the positions of documents d_i and d_j in the ranking π_{ij} (i.e., a ranking where $d_i \prec_\pi d_j$), namely $\Delta\text{rND}_{ij} = \text{rND}_{\pi_{ji}} - \text{rND}_{\pi_{ij}}$.

Note that for document pairs that either belong to the same group or appear in the same bin of the ranking, their swap does not affect the rND metric. In such cases, $\Delta\text{rND}_{ij} = 0$, which implies $\lambda_{ij}^{\text{rND}} = 0$. For efficiency, we discard these pairs when estimating $\lambda(F)_i^{\text{rND}}$.

ΔrND: Variation on swap. Given a ranking π for a query q, produced by the model at the current iteration, and let d_i and d_j be two documents in D such that d_i is ranked higher than d_j in π, i.e., $d_i \prec_\pi d_j$. If ranking d_i higher than d_j provides better fairness, then after the swap, the rND metric gets worse, and its value increases, resulting in $\Delta\text{rND}_{ij} > 0$. Conversely, $\Delta\text{rND}_{ij} < 0$ when d_j should be ranked higher than d_i, and thus the swap entails better fairness.

Finally, given any two documents d_i and d_j in D, where $d_i \prec_\pi d_j$, and their ids i and

iNEST Interconnected Nord-Est Innovation Ecosystem
Via VIII Febbraio 1848, 2 - 35122, Padova
CF 92315730280 | **Cap.Soc.** Euro 100.000,00 i.v.
Email: info@consorzioinest.it
PEC: consorzio_inest@pec.it

j, we define F as follows:

$$F = \{(i,j) \mid \Delta\text{rND}_{ij} > 0\} \cup \{(j,i) \mid \Delta\text{rND}_{ij} < 0\} \ .$$

The ordered pair (i,j) is included in F if the relative ranking $d_i \prec_\pi d_j$ must be maintained to avoid reducing rND. Conversely, (j,i) is included in F if swapping d_i and d_j in π improves rND. It is worth noting that F is created independently of E. Consequently, no optimal agreements between E and F are guaranteed.

rND+: Fairness driven. Let $\pi_{\text{rND+}}$ be a total ordering of D that maximizes NDCG while ensuring minimal rND. We then define F as follows:

$$F = \{(i,j) \mid ind_b(d_i|\pi_{\text{rND+}}) < ind_b(d_j|\pi_{\text{rND+}})\} \ ,$$

where ind_b is a function that returns the ranking-bin index containing a given document. More formally, given any ranking π, $h = ind_b(d_i|\pi)$, with $1 \leq h \leq \lceil n/b \rceil$, is the index of the bin that includes document d_i, i.e., the index h such that $d_i \in \pi[(h-1)b+1, hb]$. Thus, this definition of F reflects the bin-wise order of the ideal ranking $\pi_{\text{rND+}}$, which gives priority to fairness over effectiveness.

The ranking $\pi_{\text{rND+}}$ is constructed in two stages. The first stage is precomputed before training to improve efficiency, and produces a static ranking $\tilde{\pi}_{\text{rND+}}$ as follows. First, documents in D are sorted by decreasing relevance to *maximize* NDCG. Next, we apply the minimum number of rank promotions or demotions to the protected documents, without altering their relative order, so that each ranking bin contains a proportion of protected documents that closely matches $|\mathcal{G}^+|/n$. This step ensures minimal rND without completely disrupting the partial order induced by relevance labels.

The second stage is applied at each training iteration, where we slightly adjust the total ordering $\tilde{\pi}_{\text{rND+}}$. Note that $\tilde{\pi}_{\text{rND+}}$ is only one of the possible optimal orderings due to *ties*. Indeed, two documents with equal relevance and group membership can be swapped in $\tilde{\pi}_{\text{rND+}}$ without affecting NDCG or rND. Therefore, at each training step, we update $\tilde{\pi}_{\text{rND+}}$ into $\pi_{\text{rND+}}$ to align it with the current ranking π based on the model's predicted scores. More specifically, given a set of ties, a subset of documents in D sharing the same relevance and group, we reorder $\tilde{\pi}_{\text{rND+}}$ accordingly. The rationale for this approach is not to break what has been learned from the model so far by introducing unnecessary order constraints.

NDCG+: Effectiveness driven. This third strategy resembles rND+, but gives priority to NDCG. Let $\pi_{\text{NDCG+}}$ be a total ordering of documents that minimizes rND while preserving the *maximum* NDCG. The set F is therefore defined as follows, where F reflects the bin-wise order of the ideal ranking $\pi_{\text{NDCG+}}$, which emphasizes effectiveness over fairness:

$$F = \{(i,j) \mid ind_b(d_i|\pi_{\text{NDCG+}}) < ind_b(d_j|\pi_{\text{NDCG+}})\} \ .$$

As with $\pi_{\text{rND+}}$, the ranking $\pi_{\text{NDCG+}}$ is constructed in two stages. First, documents in D are sorted in decreasing order of relevance, thereby maximizing NDCG. Next, a

iNEST Interconnected Nord-Est Innovation Ecosystem

Via VIII Febbraio 1848, 2 - 35122, Padova
CF 92315730280 | **Cap.Soc.** Euro 100.000,00 i.v.
Email: info@consorzioinest.it
PEC: consorzio_inest@pec.it

minimal number of inter-bin swaps is applied between equally relevant documents belonging to different groups to balance protected and unprotected items across ranking bins and approximate $|\mathcal{G}^+|/n$. Since these documents have equal relevance, such swaps represent ties that do not affect NDCG but serve to reduce rND by improving group balance. The second stage proceeds in the same manner as used for generating $\pi_{\text{rND+}}$.

4.2.2 RESUSLTS

Figure 3 illustrates the trade-off between effectiveness and fairness, where fairness is measured as $(1 - \text{rND})\%$ (higher values indicate better fairness). Specifically, we present NDCG and rND values evaluated at cutoff $k = 15$ on the test sets of each dataset. With the exception of the MSLR-30K dataset, where ΔrND performs best, rND+ emerges as the overall top-performing variant, achieving higher fairness with only a slight reduction in effectiveness compared to LambdaMART and the other LambdaFair variants. Compared to the fair baselines, LambdaFair consistently achieved higher effectiveness with a slight decrease in fairness.

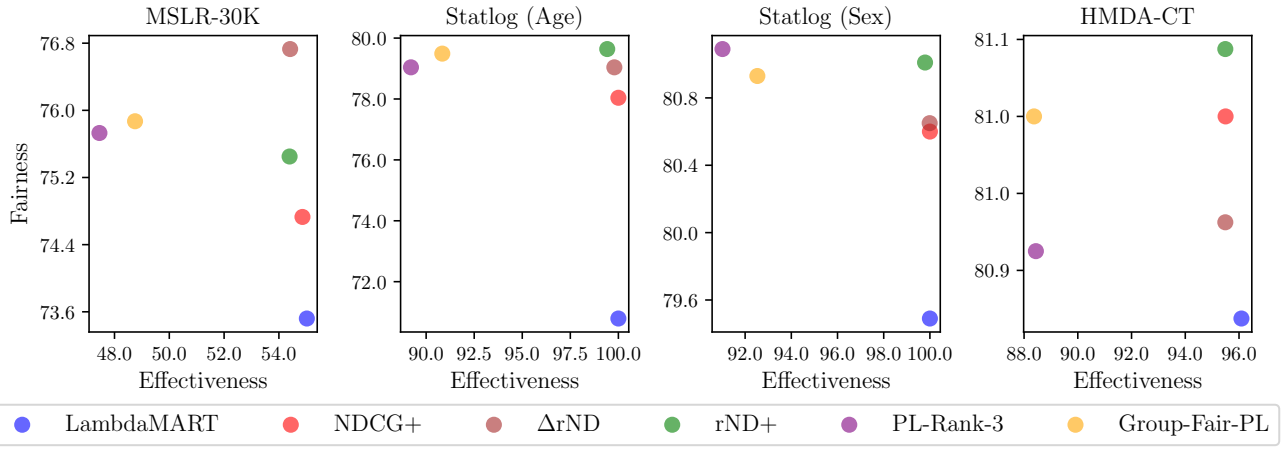

Figure 3: Effectiveness and fairness trade-off. Fairness $= (1 - \text{rND})\%$. Results for models trained and evaluated with cutoff $k = 15$.

5 CONCLUSIONS

During this research grant, I developed two novel algorithms aimed at creating fair and effective search engines: Lambda-*eX* [16,19] and LambdaFair [20,21]. Both algorithms employ an *in-processing* approach, integrating fairness constraints directly into the learning-to-rank optimization process. These strategies allow the model to simultaneously maximize ranking effectiveness while actively mitigating discrimi-

iNEST Interconnected Nord-Est Innovation Ecosystem

Via VIII Febbraio 1848, 2 - 35122, Padova
CF 92315730280 | **Cap.Soc**. Euro 100.000,00 i.v.
Email: info@consorzioinest.it
PEC: consorzio_inest@pec.it

natory biases that may arise among equally relevant items or groups requiring protection.

More specifically, these algorithms address two critical aspects of fairness in ranking systems. First, they ensure *individual fairness* by treating items with comparable relevance scores equitably during the training phase, thus reducing unwarranted discrimination between similarly qualified entities. Second, they enforce *group fairness* by promoting statistically fair exposure for protected groups across different ranking positions. This dual focus supports the development of ranking models that balance user relevance needs with fairness objectives, avoiding unfair underrepresentation of protected or disadvantaged groups.

In practical terms, these advancements have significant implications for domains such as tourism, where search results influence user decisions and economic opportunities. For instance, these algorithms can ensure that tourism-related entities, such as accommodations, attractions, or activities, that are equally relevant to a user's query receive equal consideration, preventing biases that might favor popular or centrally located options unfairly. Additionally, protected entities, such as activities located in less frequented or marginalized areas, are guaranteed proportional exposure in the ranking results. This leads to more equitable visibility and can help support diverse and sustainable tourism development.

Overall, the development and evaluation of Lambda-e*X* and LambdaFair contribute to advancing fairness-aware learning-to-rank research by demonstrating that fairness constraints can be effectively incorporated into ranking optimization without significantly compromising relevance. Future work will focus on extending these methods to handle multiple protected attributes simultaneously, exploring adaptive fairness constraints based on user preferences, and applying the algorithms to other domains where fairness in ranking is paramount.

REFERENCES

[1] A. Bower, H. Eftekhari, M. Yurochkin, and Y. Sun, "Individually fair rankings," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=71zCSP_HuBN

[2] S. Bruch, "An alternative cross entropy loss for learning-to-rank," in *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, J. Leskovec, M. Grobelnik, M. Najork, J. Tang, and L. Zia, Eds. ACM / IW3C2, 2021, pp. 118–126. [Online]. Available: https://doi.org/10.1145/3442381.3449794

[3] C. J. C. Burges, "From ranknet to lambdarank to lambdamart: An overview," 2010.

[4] C. J. C. Burges, R. Ragno, and Q. V. Le, "Learning to rank with nonsmooth cost functions," in *Advances in Neural Information Processing Systems 19, Pro-*

Via VIII Febbraio 1848, 2 - 35122, Padova
CF 92315730280 | **Cap.Soc**. Euro 100.000,00 i.v.
Email: info@consorzioinest.it
PEC: consorzio_inest@pec.it

ceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006, B. Schölkopf, J. C. Platt, and T. Hofmann, Eds. MIT Press, 2006, pp. 193–200.

[5] C. J. C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. N. Hullender, "Learning to rank using gradient descent," in *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, ser. ACM International Conference Proceeding Series, L. D. Raedt and S. Wrobel, Eds., vol. 119. ACM, 2005, pp. 89–96.

[6] O. Chapelle and Y. Chang, "Yahoo! learning to rank challenge overview," in *Proceedings of the Yahoo! Learning to Rank Challenge, held at ICML 2010, Haifa, Israel, June 25, 2010*, ser. JMLR Proceedings, O. Chapelle, Y. Chang, and T. Liu, Eds., vol. 14. JMLR.org, 2011, pp. 1–24. [Online]. Available: http://proceedings.mlr.press/v14/chapelle11a.html

[7] M. Corporation, *LightGBM Release 3.3.3.99*, 2023.

[8] F. F. I. E. Council, "HMDA Data Publication," 2017, released due to the Home Mortgage Disclosure Act. [Online]. Available: https://www.consumerfinance.gov/data-research/hmda/historic-data/

[9] P. Donmez, K. M. Svore, and C. J. C. Burges, "On the local optimality of lambdarank," in *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, and J. Zobel, Eds. ACM, 2009, pp. 460–467. [Online]. Available: https://doi.org/10.1145/1571941.1572021

[10] R. Fisher, *The design of experiments. 1935.* Edinburgh: Oliver and Boyd, 1935.

[11] S. Gorantla, E. Bhansali, A. Deshpande, and A. Louis, "Optimizing learning-to-rank models for ex-post fair relevance," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, G. H. Yang, H. Wang, S. Han, C. Hauff, G. Zuccon, and Y. Zhang, Eds. ACM, 2024, pp. 1525–1534. [Online]. Available: https://doi.org/10.1145/3626772.3657751

[12] H. Hofmann, "Statlog (German Credit Data)," UCI Machine Learning Repository, 1994, DOI: https://doi.org/10.24432/C5NC77.

[13] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, 2002. [Online]. Available: http://doi.acm.org/10.1145/582415.582418

iNEST Interconnected Nord-Est Innovation Ecosystem

Via VIII Febbraio 1848, 2 - 35122, Padova
CF 92315730280 | **Cap.Soc**. Euro 100.000,00 i.v.
Email: info@consorzioinest.it
PEC: consorzio_inest@pec.it

[14] J. Kotary, F. Fioretto, P. V. Hentenryck, and Z. Zhu, "End-to-end learning for fair ranking systems," in *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, F. Laforest, R. Troncy, E. Simperl, D. Agarwal, A. Gionis, I. Herman, and L. Médini, Eds. ACM, 2022, pp. 3520–3530. [Online]. Available: https://doi.org/10.1145/3485447.3512247

[15] P. Lahoti, K. P. Gummadi, and G. Weikum, "ifair: Learning individually fair data representations for algorithmic decision making," in *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*. IEEE, 2019, pp. 1334–1345. [Online]. Available: https://doi.org/10.1109/ICDE.2019.00121

[16] C. Lucchese, F. Marcuzzi, and S. Orlando, "Does lambdamart do what you expect?" in *Proceedings of the 13th Italian Information Retrieval Workshop (IIR 2023), Pisa, Italy, June 8-9, 2023*, ser. CEUR Workshop Proceedings, F. M. Nardini, N. Tonellotto, G. Faggioli, and A. Ferrara, Eds., vol. 3448. CEUR-WS.org, 2023, p. 72. [Online]. Available: https://ceur-ws.org/Vol-3448/paper-16.pdf

[17] C. Lucchese, F. M. Nardini, R. Perego, S. Orlando, and S. Trani, "Selective gradient boosting for effective learning to rank," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, K. Collins-Thompson, Q. Mei, B. D. Davison, Y. Liu, and E. Yilmaz, Eds. ACM, 2018, pp. 155–164. [Online]. Available: https://doi.org/10.1145/3209978.3210048

[18] R. D. Luce, *Individual Choice Behavior: A Theoretical analysis*. New York, NY, USA: Wiley, 1959.

[19] F. Marcuzzi, C. Lucchese, and S. Orlando, "Lambdarank gradients are incoherent," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, I. Frommholz, F. Hopfgartner, M. Lee, M. Oakes, M. Lalmas, M. Zhang, and R. L. T. Santos, Eds. ACM, 2023, pp. 1777–1786. [Online]. Available: https://doi.org/10.1145/3583780.3614948

[20] ——, "Lambdafair: A fair and effective lambdamart," in *Proceedings of the 14th Italian Information Retrieval Workshop, IIR 2024, Udine, Italy, September 5-6, 2024*, E. Maddalena, S. Mizzaro, K. Roitero, and M. Viviani, Eds., 2024.

[21] ——, "Lambdafair for fair and effective ranking," in *Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part IV*, ser. Lecture Notes in Computer Science, C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, and N. Tonellotto, Eds., vol. 15575. Springer, 2025, pp. 197–213. [Online]. Available: https://doi.org/10.1007/978-3-031-88717-8_15

iNEST Interconnected Nord-Est Innovation Ecosystem

Via VIII Febbraio 1848, 2 - 35122, Padova
CF 92315730280 | **Cap.Soc**. Euro 100.000,00 i.v.
Email: info@consorzioinest.it
PEC: consorzio_inest@pec.it

[22] H. Oosterhuis, "Computationally efficient optimization of plackett-luce ranking models for relevance and fairness," in *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, and T. Sakai, Eds. ACM, 2021, pp. 1023–1032. [Online]. Available: https://doi.org/10.1145/3404835.3462830

[23] ——, "Learning-to-rank at the speed of sampling: Plackett-luce gradient estimation with minimal computational complexity," in *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, and G. Kazai, Eds. ACM, 2022, pp. 2266–2271. [Online]. Available: https://doi.org/10.1145/3477495.3531842

[24] R. L. Plackett, "The analysis of permutations," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 24, no. 2, pp. 193–202, 1975. [Online]. Available: http://www.jstor.org/stable/2346567

[25] T. Qin and T. Liu, "Introducing LETOR 4.0 datasets," *CoRR*, vol. abs/1306.2597, 2013. [Online]. Available: http://arxiv.org/abs/1306.2597

[26] A. Singh and T. Joachims, "Policy learning for fairness in ranking," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 5427–5437. [Online]. Available: https://dl.acm.org/doi/10.5555/3454287.3454774

[27] A. Vardasbi, F. Sarvi, and M. de Rijke, "Probabilistic permutation graph search: Black-box optimization for fairness in ranking," in *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, and G. Kazai, Eds. ACM, 2022, pp. 715–725. [Online]. Available: https://doi.org/10.1145/3477495.3532045

[28] X. Wang, C. Li, N. Golbandi, M. Bendersky, and M. Najork, "The lambdaloss framework for ranking metric optimization," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, A. Cuzzocrea, J. Allan, N. W. Paton, D. Srivastava, R. Agrawal, A. Z. Broder, M. J. Zaki, K. S. Candan, A. Labrinidis, A. Schuster, and H. Wang, Eds. ACM, 2018, pp. 1313–1322.

[29] H. Yadav, Z. Du, and T. Joachims, "Policy-gradient training of fair and unbiased ranking functions," in *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval,*

iNEST Interconnected Nord-Est Innovation Ecosystem

Via VIII Febbraio 1848, 2 - 35122, Padova
CF 92315730280 | **Cap.Soc**. Euro 100.000,00 i.v.
Email: info@consorzioinest.it
PEC: consorzio_inest@pec.it

Virtual Event, Canada, July 11-15, 2021, F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, and T. Sakai, Eds. ACM, 2021, pp. 1044–1053. [Online]. Available: https://doi.org/10.1145/3404835.3462953

[30] K. Yang and J. Stoyanovich, "Measuring fairness in ranked outputs," in *Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA, June 27-29, 2017*. ACM, 2017, pp. 22:1–22:6. [Online]. Available: https://doi.org/10.1145/3085504.3085526

[31] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates, "Fa*ir: A fair top-k ranking algorithm," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, E. Lim, M. Winslett, M. Sanderson, A. W. Fu, J. Sun, J. S. Culpepper, E. Lo, J. C. Ho, D. Donato, R. Agrawal, Y. Zheng, C. Castillo, A. Sun, V. S. Tseng, and C. Li, Eds. ACM, 2017, pp. 1569–1578. [Online]. Available: https://doi.org/10.1145/3132847.3132938

[32] M. Zehlike and C. Castillo, "Reducing disparate exposure in ranking: A learning to rank approach," in *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, Y. Huang, I. King, T. Liu, and M. van Steen, Eds. ACM / IW3C2, 2020, pp. 2849–2855. [Online]. Available: https://doi.org/10.1145/3366424.3380048

[33] M. Zehlike, P. Hacker, and E. Wiedemann, "Matching code and law: achieving algorithmic fairness with optimal transport," *Data Min. Knowl. Discov.*, vol. 34, no. 1, pp. 163–200, 2020. [Online]. Available: https://doi.org/10.1007/s10618-019-00658-8

[34] M. Zehlike, K. Yang, and J. Stoyanovich, "Fairness in ranking, part I: score-based ranking," *ACM Comput. Surv.*, vol. 55, no. 6, pp. 118:1–118:36, 2023. [Online]. Available: https://doi.org/10.1145/3533379

[35] ——, "Fairness in ranking, part II: learning-to-rank and recommender systems," *ACM Comput. Surv.*, vol. 55, no. 6, pp. 117:1–117:41, 2023. [Online]. Available: https://doi.org/10.1145/3533380

[36] R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, ser. JMLR Workshop and Conference Proceedings, vol. 28. JMLR.org, 2013, pp. 325–333. [Online]. Available: http://proceedings.mlr.press/v28/zemel13.html