Fairness in Ranking under Disparate Uncertainty

Richa Rastogi
Cornell University
United States of America
rr568@cornell.edu

Thorsten Joachims
Cornell University
United States of America
tj@cs.cornell.edu

Abstract

Ranking is a ubiquitous method for focusing the attention of human evaluators on a manageable subset of options. Its use as part of human decision-making processes ranges from surfacing potentially relevant products on an e-commerce site to prioritizing college applications for human review. While ranking can make human evaluation more effective by focusing attention on the most promising options, we argue that it can introduce unfairness if the uncertainty of the underlying relevance model differs between groups of options. Unfortunately, such disparity in uncertainty appears widespread, often to the detriment of minority groups for which relevance estimates can have higher uncertainty due to a lack of data or appropriate features. To address this fairness issue, we propose Equal-Opportunity Ranking (EOR) as a new fairness criterion for ranking and show that it corresponds to a group-wise fair lottery among the relevant options even in the presence of disparate uncertainty. EOR optimizes for an even cost burden on all groups, unlike the conventional *Probability Ranking Principle*, and is fundamentally different from existing notions of fairness in rankings, such as *demographic parity* and *proportional Rooney rule* constraints that are motivated by proportional representation relative to group size. To make EOR ranking practical, we present an efficient algorithm for computing it in time $O(n\log(n))$ and prove its close approximation guarantee to the globally optimal solution. In a comprehensive empirical evaluation on synthetic data, a US Census dataset, and a real-world audit of Amazon search queries, we find that the algorithm reliably guarantees EOR fairness while providing effective rankings.

CCS Concepts

• **Information systems → Rankings; Top-k retrieval; Recommender systems; Decision support systems**.

Keywords

ranking, fairness, disparate uncertainty, cost of opportunity

ACM Reference Format:
Richa Rastogi and Thorsten Joachims. 2024. Fairness in Ranking under Disparate Uncertainty. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '24), October 29–31, 2024, San Luis Potosi, Mexico.* ACM, New York, NY, USA, 31 pages. https://doi.org/10.1145/3689904.3694703

EAAMO '24, October 29–31, 2024, San Luis Potosi, Mexico

ACM ISBN 979-8-4007-1222-7/24/10
https://doi.org/10.1145/3689904.3694703

1 Introduction

Human decision-processes are increasingly augmented with algorithmic decision-support systems, which has created opportunities and challenges for addressing group-based disparities in decision outcomes [5, 15, 51, 56]. In this paper, we focus on selection processes where humans evaluators use rankings to organize the order of review under resource constraints. We argue that *disparities in uncertainty* can be a major source of group-based discrimination in this setting.

To illustrate the problem, consider the following example of college admissions at a highly selective institution. In this situation, there are far more qualified candidates than available spots. Under a fixed reviewing budget, the college could give all applications a brief review (but risk high error rates in human decision making), or use a ranking to focus reviewing efforts on the more promising applications. The latter is likely to decrease error rates in human review, but it risks that this prioritization unfairly favors some groups over others. For example, consider 12,000 applicants competing for 500 slots. In this example, 10,000 applicants are from a majority group with plenty of available data, and the model can quite accurately predict which students will be admitted by the human reviewers. In particular, it accurately assigns a probability of 0.9 to 1000 of the students, and 0.01 to the remaining 9,000. The remaining 2000 applicants are from a minority group, where the model is less informed about individual students and thus assigns 0.1 to everybody. When naively ranking students by this probability, the students with 0.9 from the majority group would be ranked ahead of all the students from the minority group - and the class will fill up with the expected 900 (1000×0.9) qualified majority students before the admission staff even gets to any of the minority students. This is clearly unfair even if the predictions are perfectly calibrated for each group, since not even a single student of the expected 200 (2000×0.1) qualified students in the minority group has a chance to be selected by the admissions staff.

We aim to define a new way of ranking that does not introduce unfairness into a human decision-making process even if the predictive model shows differential uncertainty between groups. This goal recognizes that training models to have equal uncertainty across groups may be difficult in practice, since a lack of data and appropriate features for some groups may be difficult to overcome[1]. Importantly, a key principle behind our work is to leave the final decisions to human decision makers. We thus aim to design new ranking algorithms to most effectively support a fair human decision-making process, and not to replace the human decision maker.

The main contributions of this paper are

[1] Arguably, the same applies to instructing human evaluators to provide such ranking scores during a first phase of review.

- A **new fairness criterion** that provides a meaningful guarantee for rankings that are used to support human decision making in selection processes even **under disparities in uncertainty**. We motivate this fairness criterion with a fair lottery [22, 44], ensuring group-wise outcomes that are equivalent to allocating scarce resources based on a group-fair lottery among the relevant candidates.
- Based on this notion of fairness, we develop a new ranking procedure that is group-fair under disparate uncertainty. Motivated by its relation to the equality of opportunity framework [23], we name this **ranking procedure Equal Opportunity Ranking (EOR)**. We analyze EOR from the lens of the cost burden on each entity involved – the principal decision maker and each of the candidate groups – and formulate the cost to each entity as the lost opportunity of access given that the candidate was truly relevant. We show that this EOR procedure equalizes the cost burden between groups and present an efficient and practical algorithm for computing EOR rankings. This procedure always produces a near optimal and approximately EOR-fair solution. In particular, we prove an **approximation guarantee** showing that the gap in total cost to the principal compared to an optimal algorithm is bounded by a small amount.
- In addition to these theoretical worst-case guarantees, we present **extensive experiments** benchmarking the EOR algorithm with various existing ranking algorithms under different settings of disparate uncertainty. We show that Demographic Parity [58, 61], normative procedures like Proportional Rooney-rule-like constraints [9], Exposure based fairness criteria [49], and Thompson Sampling Policy [50] are not typically EOR-fair under disparate uncertainty. We find that these results hold on both a wide range of synthetic datasets, as well as on real-world US census data. Finally, we explore the use of our fairness criterion for auditing ranking systems, using a real-world dataset of Amazon shopping search queries. Our code can be accessed at https://github.com/RichRast/DisparateUncertainty.

These results have important societal implications. First, they provide evidence that naively applying existing fairness mechanisms in rankings under disparate uncertainty leads to unfairness in terms of one group bearing the majority of the cost of opportunity. Second, even under high disparate uncertainty in the worst case, EOR guarantees an approximately equal cost burden among all groups with bounded additional cost to the human decision maker. Finally, we hope our results inform practitioners to collect data and appropriate features for candidates in all groups to build predictive models that reduce disparate uncertainty. As we will show, the EOR procedure elevates the candidates with high uncertainty in the rankings for human evaluation. This has the desirable effect of producing more equitable training data for future use.

We now highlight some important considerations here. First, our proposed method is grounded in the fairness of a lottery [45], which is a common technique for allocating scarce resources (e.g., admission slots among a large number of qualified candidates). However moral and philosophical arguments debating the use of lottery and randomization for certain situations have also been made [26]. We hope this work can spark discussions on alternative notions of fairness in rankings that satisfy equality of opportunity under disparate uncertainty. Another important point is that our proposed EOR procedure reduces unfairness due to disparate uncertainty, which often but not necessarily coincides with the historically disadvantaged group. Since EOR doesn't require the designation of the disadvantaged group, the guarantees we provide are not making a normative statement about any historically disadvantaged group. To that end, we emphasize the careful consideration of historical and social context that needs to be taken into account by the human decision maker as well as the way groups are defined in the first place.

2 Related Works

While the issue of fairness has been heavily studied in the classification setting, its counterpart – the ranking setting has received relatively less attention. Below we highlight key areas related to our work and leave a more detailed discussion of these and other related works to Appendix B.

Fairness in Rankings and Selection Processes: While there exist several notions of fairness in rankings [64], predominantly, they are variations of two fairness mechanisms in existing literature – representation by size [10, 57, 61, 63] and equitable allocation of exposure [4, 31, 35, 48, 49]. We propose a new criterion different from either of the two and our central point is that under disparate uncertainty between groups, it is more fair to take an equal proportion of relevance in expectation rather than equality by size or exposure. Proportional representation in the form of diversity constraints like demographic parity [58] or affirmative action such as the Rooney Rule [9] guarantee a minimum proportion by group size in selection processes. Exposure based formulations in rankings ensure that groups of candidates are allocated exposure in an equitable way such as in proportion of amortized relevance over the full ranking [4]. In this work, we demonstrate that fairness of representation by size and exposure, are not sufficient under disparate uncertainty.

Fairness in Rankings under Uncertainty: Our work builds on [50], in which the authors establish that uncertainty in relevance probabilities is a primary cause of unfairness for rankings. They propose a Thompson sampling policy that randomizes relevances drawn from the predictive posterior distribution. Separately, [19] studies the role of affirmative action in the presence of differential variance between groups in rankings. Differential variance implies that there is more certainty about the true quality (scores) of candidates in a group with less variance in the estimated quality and vice versa for a group with higher variance. In contrast, we work with relevance probabilities instead of scores and focus on the certainty of relevance of a candidate, which is determined by how close the predicted relevance probabilities are to 1 or 0. For instance, a group is highly certain (if the probabilities are all close to 1.0) or highly uncertain (if the probabilities are all close to 0.5) while both groups could have similar variance in probabilities. Fairness under uncertainty has also been studied with respect to calibration of probabilities [11, 20, 29, 38]. Classical literature in this area studies whether group-wise calibration is a necessary condition for fairness, or not [32]. Our work is orthogonal to the question of the necessity of calibration for fairness and we only require group-wise calibration as a sufficient condition for the EOR criterion we propose.

Our work complements and extends prior research on fairness in rankings under uncertainty, contributing uniquely in several ways. In particular, we provide a formal framework for analyzing the unfairness that differential uncertainty induces in rankings. Additionally, our approach involves accounting for the differential uncertainty directly at the ranking stage, unlike prior work that involves learning the uncertainty [53] or correcting the noisy relevance estimates [59]. Finally, our proposed EOR criterion is non-amortized for every prefix k of the ranking, which is strictly stronger than the probabilistic but amortized notions of fairness [4, 48, 49] shown to be problematic [28].

3 Un-fairness due to Disparate Uncertainty in Rankings

We want to design a ranking policy π that does not introduce unfairness into a human decision process due to disparate uncertainty. More formally, the task of π is to compute a ranking σ of n candidates, where each candidate i has a binary[2] relevance $r_i \in \{0, 1\}$ which is unknown to the ranking policy π, and true relevance can only be revealed through a human decision maker. When assessing the relevance, we assume that the human decision maker goes through the ranking σ from the top to some a priori unknown position k. The goal of the decision maker (a.k.a. principal) is to find as many relevant candidates (e.g., relevant products, qualified students) as possible.

While the true relevances r_i are unknown, we assume that the ranking policy π has access to a predictive model of relevance $\mathbb{P}(r_i|\mathcal{D})$, typically trained on prior human decisions $\mathcal{D}$ and features of the candidates. Sorting the candidates in decreasing order of $p_i = \mathbb{P}(r_i = 1|\mathcal{D})$ is called the Probability Ranking Principle (PRP) [41], and it is by far the most common way of computing a ranking. The justification for PRP ranking is that it maximizes the expected number of relevant candidates in any top-k prefix of the ranking. On the other hand, Demographic Parity (DP) is the dominant form of fairness mechanism in rankings, where candidates are selected from groups in proportion to the group size. While PRP ranking is provably optimal according to the efficiency goal of the principal and DP ranking ensures representation by group size, the following elaborates how both PRP and DP can violate fairness.

3.1 Illustrative Example

Consider a medical setting, where candidates need to be evaluated for eligibility to participate in a controlled medical trial. While group A consists of candidates with a rich set of diagnostic tests that inform eligibility (e.g., candidates with health insurance), group B consists of candidates without prior access to such tests (e.g., candidates without health insurance). As a result, according to $\mathbb{P}(r_i = 1|\mathcal{D})$ in Figure 1, the model can make very informed predictions for candidates in group A, while for group B the model cannot reliably differentiate between eligible and not eligible candidates. This means the model knows exactly which candidates in group A will be judged as eligible by the human decision maker, but it will make undifferentiated (but well-calibrated) predictions for candidates in group B.

Figure 2 shows that the PRP ranking is oblivious to this disparity between groups. If the principal needs to find four eligible candidates based on the PRP ranking, they are all selected from group A. However, by summing the probabilities in group B, our model tells us that we can also expect four eligible candidates in group B. We argue that deterministically selecting only candidates from group A is unfair since it is not consistent with the outcome of a group-fair lottery for the four spots among the eight eligible candidates. Now, consider the DP ranking in Figure 2. Since group A has 17 candidates and group B has 8 candidates, DP will select roughly one candidate from group B for every two candidates from group A. We argue that in this setting, DP is also unfair, (though less in comparison to PRP) as it selects three eligible candidates from group A and only one from group B. In expectation, it selects 2.6 out of 4 relevant candidates from group A, but only 0.6 out of 4 relevant candidates from group B. We show empirically later that other fairness mechanisms motivated by representation of size such as proportional Rooney Rule or threshold-based formulations have the same failure mode. Importantly, note that it is not evident whether group A or B should be the majority group.

We argue that a more principled and fair way would be to select an equal fraction of relevant candidates from each group in expectation. Consider the last ranking in Figure 2, which approximately fulfills the EOR fairness we formally introduce later. In expectation, this ranking selects a more equal number of relevant candidates from both groups, making it similar to a fair lottery. In particular, it selects 1.8 out of 4 relevant candidates from group A and 1.2 out of 4 relevant candidates from group B. This EOR ranking, however, comes at an increased evaluation cost to the principal as it selects 3.0 expected relevant candidates from both the groups, compared to 3.2 with DP and 3.3 with PRP. As a result, the principal needs to review more candidates to select the same number of relevant candidates with EOR ranking. However, it is still far more effective than a lottery, which selects the candidates in a uniform random order.

Our key insight is that EOR ranking is more fair not because it takes an equal "number" of candidates from each group but it is more fair because it takes an equal fraction of "relevant" candidates in expectation from each group. This accounts for predictive uncertainty in the relevance probabilities because even when one group has sharp and the other group has non-sharp p_i, it takes approximately equal fraction of relevance from each of the groups.

This example illustrates the intuition behind the EOR principle we formalize in the following, and we will show how to efficiently compute rankings that fulfill EOR fairness.

3.2 Sources of Disparate Uncertainty

It remains to show that disparate uncertainty is a fundamental problem when estimating the relevance probabilities $\mathbb{P}(r_i = 1|\mathcal{D})$ that is not easily remedied by improved learning methods. The following illustrates that even a Bayes-optimal procedure is vulnerable to producing disparate uncertainty.

Consider the posterior distribution illustrated in Figure 3, which shows the uncertainty $\mathbb{P}(\theta_i|\mathcal{D})$ that a Bayesian model has about

[2] We conjecture that our framework can be extended to categorical or real-valued relevances.

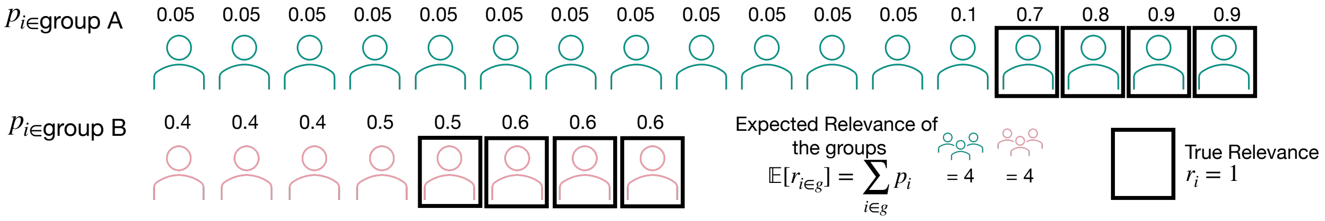


Figure 1: The expected probability of relevance p_i and their true relevance r_i for all candidates in both groups.

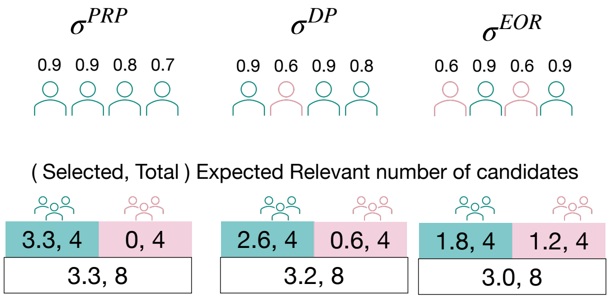


Figure 2: Top-4 ranking for Probability Ranking Principle (PRP), Demographic Parity (DP), and our proposed EOR for the example in Figure 1. Selected relevant number of candidates in expectation and total relevant number of candidates in expectation are shown corresponding to each ranking.

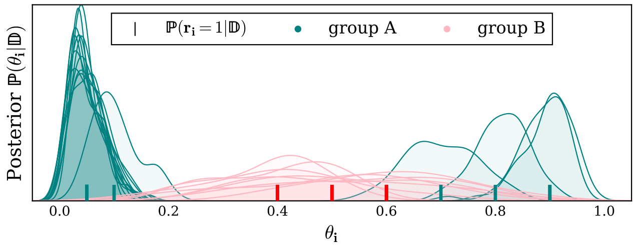


Figure 3: An illustration of disparate uncertainty between groups from a Bayesian perspective for all the candidates of Figure 1. The candidates in group A have peaky posteriors, while those in group B have relatively flat posteriors.

the relevance probability θ_i of candidate i, where θ_i is the parameter of a Bernoulli distribution. For group A, the posterior $\mathbb{P}(\theta_i|\mathcal{D})$ is peaked, meaning that the model can accurately pinpoint the correct relevance probabilities. For group B, the posterior is flat, which is to be expected if group B is smaller and thus has less data. The Bayes-optimal way of handling this uncertainty is to infer $\mathbb{P}(r_i|\mathcal{D})$ via the posterior predictive distribution

$$\mathbb{P}(r_i = 1|\mathcal{D}) = \int \mathbb{P}(r_i = 1|\theta_i)\,\mathbb{P}(\theta_i|\mathcal{D})\,d\theta_i = \int \theta_i\,\mathbb{P}(\theta_i|\mathcal{D})\,d\theta_i$$

Figure 3 shows how even this Bayes-optimal procedure leads to disparate uncertainty between groups, where the $\mathbb{P}(r_i = 1|\mathcal{D})$ is closer to zero or one for candidates in group A (i.e., highly informative), and middling for group B (i.e., less informative).

Note that there is ample evidence that non-Bayesian methods also produce such disparities (e.g., [5, 51, 56]). Furthermore, disparate amounts of data are not the only cause for disparity. For example, in college admissions, disparately more URM candidates may miss AP grades because their school does not offer AP classes. Their epistemic uncertainty [27] of qualification will thus be higher since the model has less information about these students. This higher uncertainty does not mean individual students are not qualified, and elevating them in the ranking for human evaluation can accurately reveal qualification through additional information (e.g., an interview, deep reading of the SOP, or recommendation letters). But if they are never selected for human review, then they do not have a chance for an admission spot.

4 Equality of Opportunity in Ranking

In this section, we first discuss the assumptions and modeling choices and then formulate the cost that the uncertainty of the predictive model imposes on the principal and the relevant candidates from the different groups.

Our first assumption includes access to group-wise calibration [3, 38] with the probability estimates calibrated within groups. To simplify notation, we do not differentiate between $\mathbb{P}(r_i|\mathcal{D})$ and a group-wise calibrated score $\mathbb{P}(r_i|s, A, \mathcal{D}) = s$ and we only require this group-wise calibration as a sufficient condition for our framework. Additionally, we assume that the true relevance r_i is revealed perfectly to the human decision-maker upon review, and we do not model any bias in the human decision-making review process. Finally, we assume that candidates have group membership to a single protected attribute and do not consider intersectional group membership, which is a practically important consideration in fairness. Relaxing these three assumptions for future work could allow modeling even more real-world complexities.

To formulate the cost of opportunity, we first recognize that any group-wise calibrated model allows us to compute the **expected number of relevant candidates** $nRel(.)$ **of a particular group** g – no matter how well the model can differentiate relevant and non-relevant candidates in that group.

$$nRel(g) = \sum_{i\in g} \mathbb{E}_{r_i\sim\mathbb{P}(r_i|\mathcal{D})}[r_i] = \sum_{i\in g} \mathbb{P}(r_i = 1|\mathcal{D})$$

Extending this to rankings, the expected number of relevant candidates from group g for any prefix k of ranking σ that only depends on $\mathbb{P}(r_i = 1|\mathcal{D})$ to ensure unconfoundedness is

$$nRel(g|\sigma_k) = \sum_{i \in g \cap \sigma_k} \mathbb{E}[r_i] = \sum_{i \in g \cap \sigma_k} \mathbb{P}(r_i = 1|\mathcal{D})$$

Further extending this to a potentially stochastic ranking policy π that represents a distribution over rankings for a particular query leads to

$$\begin{aligned} nRel(g|\pi_k) &= \sum_{i \in g} \mathbb{E}_{r_i \sim \mathbb{P}(r_i|\mathcal{D}), \sigma_k \sim \pi}[r_i \mathbb{1}_{i \in \sigma_k}] \\ &= \sum_{i \in g} \mathbb{P}(i \in \sigma_k^{\pi}) \mathbb{P}(r_i = 1|\mathcal{D}) \end{aligned} \tag{1}$$

where $\mathbb{P}(i \in \sigma_k^{\pi}) = \mathbb{E}_{\sigma_k \sim \pi}[\mathbb{1}_{i \in \sigma_k}]$ is the probability that policy π ranks candidate i into the top k. As a side note notation-wise, for a specific policy, for example, π^{EOR}, we denote the corresponding ranking $\sigma_k^{\pi^{EOR}}$ in the abbreviated form as σ_k^{EOR}.

The ability to compute these expected numbers of relevant candidates from each group allows us to reason about the cost resulting from the uncertainty of the model that each ranking imposes on the respective groups, which we detail in the following.

4.1 Cost Burden to Candidate Groups and the Principal

We define the **cost** $c(.)$ **to candidate** i as missing out on the opportunity to be selected if the candidate was truly relevant. For a ranking policy π that produces rankings $\sigma \sim \pi$ based on $\mathbb{P}(r_i|\mathcal{D})$, and a principal that reviews the top k candidates, the cost to a relevant candidate i is the probability of not being included in the top k.

$$c(i|\pi_k, r_i) = r_i(1 - \mathbb{P}(i \in \sigma_k^{\pi})) \tag{2}$$

Note that only relevant candidates can incur a cost, since non-relevant candidates will be rejected by human review and thus draw no utility independent of whether they are ranked into the top k. Also, note that $\mathbb{P}(i \in \sigma_k^{\pi})$ can be estimated by Monte-Carlo sampling even for complicated ranking policies that have no closed-form distribution.

While determining the cost to a specific individual i is difficult since it involves knowledge of the true relevance r_i, getting a measure of the aggregate cost to the group is more tractable. In particular, we define the **group cost** as the expected cost to the relevant candidates in the group, normalized by the expected number of relevant candidates.

$$\begin{aligned} c(g|\pi_k) &= \frac{\sum_{i \in g} \mathbb{E}_{r_i \sim \mathbb{P}(r_i|\mathcal{D})}[c(i|\pi_k, r_i = 1)]}{nRel(g)} \\ &= \frac{\sum_{i \in g}(1 - \mathbb{P}(i \in \sigma_k^{\pi}))\mathbb{P}(r_i = 1|\mathcal{D})}{nRel(g)} \\ &= 1 - \frac{nRel(g|\pi_k)}{nRel(g)} \end{aligned} \tag{3}$$

The last equality in (3) follows directly from Eq. (1). We normalize the expected group cost with the total expected number of relevant candidates in the group so that the above approximates the fraction of relevant candidates from that group that miss out on the opportunity of being selected by the human reviewers.

The **principal incurs a cost** whenever the ranking misses a relevant candidate, independent of group membership. For a principal that reviews the top k applications from two groups – A and B, the **total cost** can thus be quantified via the expected number of relevant candidates that are overlooked.

$$c(\text{Principal}|\pi_k) = \frac{\sum_i (1 - \mathbb{P}(i \in \sigma_k^{\pi}))\mathbb{P}(r_i = 1|\mathcal{D})}{nRel(A) + nRel(B)} \tag{4}$$

We again normalize this quantity to make it proportional to the total expected number of relevant candidates. Note that Eq. (4) is related to the conventional metric of Recall@k.

4.2 Equality of Opportunity Ranking (EOR) Criterion

We now formally define our EOR fairness criterion and argue that a disparity in uncertainty should not lead to disparate costs for any of the groups. We have already seen that π^{PRP} and π^{DP} can violate this goal. For a possible solution, we turn to the principle of random lottery that has been historically used to justify fair allocation of resources [22, 44]. Take, for example, the uniform ranking policy π^{unif}, which ignores $\mathbb{P}(r_i|\mathcal{D})$ and picks a ranking uniformly at random. Use of π^{unif} ensures that any relevant candidate has an equal chance of being evaluated and selected since any top k of the ranking contains a uniform random sample of the *relevant* candidates – independent of group membership. While the ranking effectiveness of π^{unif} is bad, it has the attractive property that the fraction of relevant candidates that get selected from each group is equal in expectation. For example, if both group A and group B contain 100 relevant candidates in expectation and if π^{unif} selects l relevant candidate in expectation from group A, it also selects l relevant candidates in expectation from group B. Similarly, if group A contains 200 relevant candidates and group B contains 100, the selection ratio will be 2 to 1 in expectation. We formalize this property of the uniform lottery as our key fairness axiom.

Axiom 1 (EOR Fair Ranking Policy). *For two groups of candidates A and B, a ranking policy π is Equality-of-Opportunity fair, if for every k the top-k subsets π_k contain in expectation an equal fraction of the relevant candidates from each group. More precisely:*

$$\forall k \quad \frac{nRel(A|\pi_k)}{nRel(A)} = \frac{nRel(B|\pi_k)}{nRel(B)} \tag{5}$$

While this fairness property of π^{unif} is desirable, its completely uninformed rankings come at a cost to the principal and the relevant candidates from both groups, since only a few relevant candidates will be found. The uniform policy π^{unif} is particularly inefficient when the fraction of relevant candidates is small. The key question is thus whether we can define an alternate ranking policy that retains the group-wise fairness properties of π^{unif}, but retains as much effectiveness in surfacing relevant candidates as possible.

To illustrate that such rankings exist, which are both EOR fair and more effective, consider our motivating example of Figure 1, where $\sigma^{EOR} = [\overset{B}{0.6}, \overset{A}{0.9}, \overset{B}{0.6}, \overset{A}{0.9}, \overset{B}{0.6}, \overset{B}{0.5}, \overset{A}{0.8}, \overset{B}{0.5}, \overset{A}{0.7}, \overset{B}{0.4}, \overset{A}{0.1}, \overset{B}{0.4}, \overset{A}{0.05}, \overset{A}{0.05}, \overset{A}{0.05}, \overset{A}{0.05}, \overset{A}{0.05}, \overset{A}{0.05}, \overset{A}{0.05}, \overset{A}{0.05}, \overset{B}{0.4}, \overset{A}{0.05}, \overset{A}{0.05}, \overset{A}{0.05}, \overset{A}{0.05}]$ has the property that the expected number of relevant candidates for each group in the top k never differs by more than 0.6 for any value of k. In one way, this guarantee is even stronger than what is defined in Axiom 1, since it holds for the specific ranking σ^{EOR} without the need for stochasticity in the ranking policy. This provides a non-amortized notion of fairness, which is particularly desirable for high-stakes ranking tasks that do not repeat, and we thus need to provide the strongest possible guarantees for the specific ranking σ we present. However, a guarantee for an individual ranking makes the problem inherently discrete, which means that we require some tolerance (i.e., 0.6 in the example above) in the fairness criterion depending on the choice of k. This leads to the following δ-EOR Fairness criterion for an individual ranking σ.

Definition 4.1 (δ-EOR Fair Ranking). For two groups of candidates A and B, a ranking σ is δ-EOR fair, if for every k the top-k subset σ_k differs in its fraction of expected relevant candidates from each group by no more than δ. More precisely:

$$\forall k \quad \left| \frac{nRel(A|\sigma_k)}{nRel(A)} - \frac{nRel(B|\sigma_k)}{nRel(B)} \right| \leq \delta \tag{6}$$

Note that we can also define a specific "slack" $\delta(\sigma_k)$ for each position k. For a fair ranking σ, this slack should ideally oscillate close to zero as we increase k, and so minimizing its deviation from zero would translate to ensuring δ-EOR fairness. Formally, we can define $\delta(\sigma_k)$ as

$$\forall k \quad \delta(\sigma_k) = \frac{\sum_{i \in A \cap \sigma_k} \mathbb{P}(r_i|\mathcal{D})}{\sum_{i \in A} \mathbb{P}(r_i|\mathcal{D})} - \frac{\sum_{i \in B \cap \sigma_k} \mathbb{P}(r_i|\mathcal{D})}{\sum_{i \in B} \mathbb{P}(r_i|\mathcal{D})} \tag{7}$$

δ-EOR fairness balances the selection of candidates from the two groups, accounting for predictive uncertainty in their estimation of relevances. If for instance, the ML model is less certain in its predictions for group B, but both groups have the same total expected relevance, the δ-EOR criterion will rank candidates from group B higher to ensure fairness. Importantly, note how this produces more human relevance labels of candidates from groups with high uncertainty, which has the desirable side-effect of producing new training data that allows training of more equitable relevance models for future use.

Finally, note how the δ-EOR fair ranking provides a means for ensuring procedural fairness and avoiding *disparate treatment*. Importantly, we leave the decision of which candidates to select to the human decision maker, and EOR fairness does not require the designation of a disadvantaged group. Instead, the EOR fair condition in Eq. (6) is symmetrical w.r.t. both groups and by definition treats both groups similarly, and its intervention in the ranking process is entirely driven by the predictive model $\mathbb{P}(r_i|\mathcal{D})$. Even though it uses group membership, EOR-fairness is thus fundamentally different from demographic parity [17, 58] and affirmative action rules like Rooney rule [9, 12], $\frac{4}{5}$th rule (selection rate for a protected group must be at least 80% of the rate for the group with the highest rate)[3] or γ-based notions of fairness [18] and threshold based formulations such as FA*IR [61].

To illustrate the difference with existing fairness notions, we return to our running example from Figure 1. For top-4 ranking in Figure 2, the EOR criterion can be computed as $|\delta(\sigma_4^{EOR})| = 0.15$, $|\delta(\sigma_4^{DP})| = 0.5$ and $|\delta(\sigma_4^{PRP})| = 0.83$, quantifying the unfairness of DP and PRP as compared to EOR. While DP selects one candidate from group B for every two candidates from group A, applying $\frac{4}{5}$th rule with group B as the disadvantaged group will select roughly 4/5 number of candidates from group B for every two candidates from group A. For top-4 ranking, the $\frac{4}{5}$th rule is $\sigma_4^{\text{FourFifth}} = [\overset{A}{0.9}, \overset{B}{0.6}, \overset{A}{0.9}, \overset{A}{0.8}]$ with $|\delta(\sigma_4^{\text{FourFifth}})| = 0.5$. If instead, group A is selected as the disadvantaged group, $\frac{4}{5}$th rule will select all four candidates from group A resulting in $|\delta(\sigma_4^{\text{FourFifth}})| = 0.83$, same as that of PRP. The FA*IR criterion (π^{FS}) is similarly anchored on the principle that a top-k ranking is fair when the proportion of disadvantaged candidates selected doesn't fall far below a required minimum proportion and also requires the designation of a disadvantaged group. In this example, π^{FS} gives the exact same top-4 ranking and EOR criterion as shown for $\frac{4}{5}$th rule. In summary, the predominant fairness criteria in rankings motivated by the representation of size perform very differently than the π^{EOR}. As an example, consider the well-documented issue of female candidates not being selected for leadership positions primarily due to their small applicant pool size [25]. If the female applicants have high disparate uncertainty (due to lack of historical data), affirmative action may still select far fewer (based on group size) of them than deserved (based on the number of relevant female candidates).

We now briefly consider two other notions of fairness in rankings for the running example. First, we look at the exposure-based formulations[4, 49]. The principle of exposure is motivated by position bias in rankings and ensures the allocation of position in rankings in proportion to the expected total relevance. While the position of a selected candidate is certainly important, it does not take disparate uncertainty into consideration. π^{EXP} is a stochastic policy that allocates equal exposure between the two groups (in this example, both groups have an equal expected total relevance) over the full 25 positions of the ranking. π^{EXP} allocates most of the probability mass to candidates in group B for all of the top-4 positions (not because they have high uncertainty but because their group size is smaller than group A). This results in a high cost burden for group A and the EOR criterion is computed as $|\delta(\sigma_4^{EXP})| = 0.58$ higher than both π^{EOR} and π^{DP}. Later in Section 7, we demonstrate how π^{EXP} places a higher cost burden on the uninformative group instead when both groups have relatively the same size.

Finally, we discuss the Thompson Sampling based fairness in rankings [50]. For π^{TS}, binary relevances are drawn according to $r_i \sim \mathbb{P}(r_i|\mathcal{D})$, and candidates are sorted in decreasing order of relevance r_i with their ranking randomized for the same value of relevance. The EOR criterion for a top-4 ranking produced by π^{TS} can be computed as $|\delta(\sigma_4^{TS})| = 0.29$ for the running example. While π^{TS} takes the predictive uncertainty of relevance into account by randomization of rankings, it is group oblivious and so does not account for the difference in the predictive uncertainty of relevance between groups. This explains the high EOR criterion of a specific σ^{TS} with median $\sum_{k=1}^{n} |\delta(\sigma_k^{TS})|$ as compared to that of the σ^{EOR}. While we discussed how EOR differs from existing fairness notions above, we will further demonstrate this comparison via extensive empirical evaluations in Section 7.

One of our key contributions includes formalizing the connection between δ-EOR Fair Ranking described in Definition 4.1 and the

[3] Uniform Guidelines on Employment Selection Procedures, 29 C.F.R.§1607.4(D) (2015)

Algorithm 1: EOR Algorithm

```
Input: Groups $g \in \{A, B\}$; Rankings $\sigma^{PRP,g}$ per group in the sorted (decreasing) order of relevance probabilities $\mathbb{P}(r_i|\mathcal{D})$.
Initialize: $j \leftarrow 0$; empty ranking $\sigma^{EOR}$
while $j < k$ do
    $l_g \leftarrow \sigma^{PRP,g}[1] \quad \forall g \in \{A, B\}$
    $g^* \leftarrow \arg\min_{g \in \{A,B\}} |\delta(\sigma^{EOR} \cup \{l_g\})|$,
    where $\delta(.)$ is computed using (7)
    $l_{g^*} \leftarrow \sigma^{PRP,g^*}[1]; \quad \sigma^{PRP,g^*} \leftarrow \sigma^{PRP,g^*} \setminus \{l_{g^*}\}$
    $\sigma^{EOR} \leftarrow \sigma^{EOR} \cup \{l_{g^*}\}; \quad j \leftarrow j + 1$
Return $\sigma^{EOR}$
```

cost of opportunity in rankings described in Section 4.1. Both δ-EOR Fair Ranking and cost of opportunity in rankings are derived separately – the former from the axiom of fairness of a uniform lottery, the latter from the cost of errors that any realistic prediction model is bound to make. In the next section, we show that these two are elegantly related via theoretical results on cost optimality.

5 Computing EOR-Fair Rankings

We now turn to the question of how to compute a δ-EOR fair ranking σ^{EOR} for any given relevance model $\mathbb{P}(r_i|\mathcal{D})$. This ranking procedure needs to account for two potentially opposing goals. First, it needs to ensure that δ-EOR fairness is not violated, ideally for a δ that is not larger than required by the discreteness of the ranking. Second, it should maximize the number of relevant candidates contained in the top k, for any a-priori unknown k. While solving this optimization problem in the exponentially sized space of rankings is computationally inefficient, we show that Algorithm 1 is an efficient ranking method that provides a close-to-optimal solution.

Algorithm 1 uses as input the PRP rankings $\sigma^{PRP,A}$ and $\sigma^{PRP,B}$ for each of the groups A and B respectively. We denote $\sigma^{PRP,g}[i]$ as the i^{th} element in the PRP ranking of group g. The basic idea is to compare the highest relevance candidate from each group and select the candidate that would minimize the δ for the resultant ranking (breaking ties arbitrarily when selecting an element from either group results in the same δ for the resultant ranking). Consider our running example from Figure 1. At $k = 1$, selecting the first element from group A, $\sigma^{PRP,A}[1]$, would result in a $\delta(\sigma_1) = 0.9/4$ while selecting the first element from group B, $\sigma^{PRP,B}[1]$, would result in a $\delta(\sigma_1) = -0.6/4$. To minimize $|\delta(\sigma_1)|$, the algorithm selects the first element from group B with $\sigma^{EOR}_1 = [\overset{B}{0.6}]$, $|\delta(\sigma_1)| = 0.6/4$. For $k = 2$, the first element from group A, and the second element from group B are considered. It proceeds to select the first element from group A with $\sigma^{EOR}_2 = [\overset{B}{0.6}, \overset{A}{0.9}]$, $|\delta(\sigma_2)| = 0.3/4$ and so on. The Algorithm does not change the relative ordering between candidates within a group and its runtime complexity is $O(n \log n)$, since the elements from the two groups each need to be sorted once by $\mathbb{P}(r_i|\mathcal{D})$. Composing the final EOR ranking σ^{EOR} by merging the two group-based rankings $\sigma^{PRP,A}$ and $\sigma^{PRP,B}$ takes only linear time since each computation per iteration is constant time per prefix k.

While Algorithm 1 is inspired by existing algorithms such as [61] in that both select the top element from the PRP ranking of each group, they are fundamentally different. Existing methods including [61] ensure a form of demographic parity which we have already shown to be fundamentally different than the EOR criterion we propose. Additionally, while [61] requires a threshold input and the designation of a disadvantaged group, the EOR Algorithm does not require this normative designation and guarantees EOR fairness without requiring any tolerance δ as an input. We show this both theoretically and in empirical evaluations and provide a detailed description of baseline algorithms in Appendix E.1.

It remains to be shown that Algorithm 1 always produces a ranking σ^{EOR} with small δ while surfacing as many relevant candidates as possible in any top k prefix. We break the proof of this guarantee into the following steps. First, we show that for any particular k and its associated $\delta(\sigma^{EOR}_k)$, the number of relevant candidates in the top-k is close to optimal. Second, we provide an upper bound on $\delta(\sigma^{EOR}_k)$ that is entirely determined a priori by the specific $\mathbb{P}(r_i|\mathcal{D})$. To address the first step, the following Theorem 5.1, shows that the rankings produced by Algorithm 1 have a cost to the principal that is close to optimal.

Theorem 5.1 (Cost Approximation Guarantee at k). *The EOR fair ranking σ^{EOR} produced by Algorithm 1 is at least $\phi\delta(\sigma^{EOR}_k)$ cost optimal for any prefix k, where $\phi = \frac{2}{nRel(A)+nRel(B)}\left|\frac{p_A - p_B}{q_A + q_B}\right|$, $q_A = \frac{p_A}{nRel(A)}$, and $q_B = \frac{p_B}{nRel(B)}$. Further, $p_A = \sigma^{PRP,A}[k_A]$, $p_B = \sigma^{PRP,B}[k_B]$, where k_A is the last element from group A that was selected by EOR Algorithm for prefix k and similarly for k_B.*

Proof Sketch: We use linear duality to prove this theorem. To find a lower bound on the cost optimal ranking that satisfies the EOR fairness constraint, we formulate the corresponding Linear Integer Problem (ILP) for selecting the optimal top-k subset under the δ-EOR constraint. This leads to the following optimization problem, where $X \in \{0, 1\}^n$ is the variable for whether the i^{th} candidate was chosen or not, P is the relevance probability for all candidates.

Minimize total cost as defined in Eq. (4)

$$\min_{x \in \{0,1\}} \quad 1 - \frac{P^T X}{nRel(A) + nRel(B)} \tag{ILP}$$

$$\text{s.t.} \quad X^T \mathbb{1} = k \qquad \text{(select up to } k \text{ candidates)}$$

$$-\delta(\sigma^{EOR}_k) \leq \left(\frac{P\mathbb{1}_A}{nRel(A)} - \frac{P\mathbb{1}_B}{nRel(B)}\right)^T X \leq \delta(\sigma^{EOR}_k)$$

$$\text{(EOR fairness from Eq. (6) must be satisfied } \forall k)$$

We relax this ILP to a Linear Program (LP) by turning any integer constraints $x \in \{0, 1\}$ in the primal into $0 \leq x \leq 1$. For the relaxed LP, we formulate its dual and construct a set of dual variables λ corresponding to the solution from the EOR Algorithm. Using the dual value of the EOR solution and the relaxed LP solution, we obtain an upper bound of the duality gap. Since the upper bound on this duality gap is w.r.t. the relaxed LP solution, it is also an upper bound for the optimal ILP solution. We provide a complete proof of the theorem and associated lemmas in Appendix C.1. □

Note that ϕ depends only on the relevance probabilities of the last elements selected from each group by the EOR Algorithm in the

k^{th} position. Furthermore, note that the solution of Algorithm 1 is the exact optimum for any k where the unfairness $\delta(\sigma_k^{EOR})$ is zero, indicating that any suboptimality of the EOR algorithm is merely due to some (presumably unavoidable) discretization effects.

While the previous theorem characterized cost optimality, the following Theorem 5.2 shows that the magnitude of unfairness $\delta(\sigma_k^{EOR})$ is bounded by some δ_{max}, providing an a priori approximation guarantee for both the amount of unfairness and the cost optimality of Algorithm 1.

Theorem 5.2 (Global Cost and Fairness Guarantee). *Algorithm 1 always produces a ranking σ^{EOR} that is at least $\phi\delta_{max}$ cost optimal for any k, with $\delta_{max} = \frac{1}{2}\left(\frac{\sigma^{PRP,A}[1]}{nRel(A)} + \frac{\sigma^{PRP,B}[1]}{nRel(B)}\right)$.*

Proof Sketch: We show via an inductive argument that according to the EOR algorithm, minimizing $\left|\delta(\sigma_k^{EOR})\right|$ at every k ensures that the resultant EOR ranking always satisfies $\delta(\sigma_k^{EOR}) \leq \frac{1}{2}\left(\frac{\sigma^{PRP,A}[1]}{nRel(A)} + \frac{\sigma^{PRP,B}[1]}{nRel(B)}\right)$, that is bounded by the average of the relevance proportions from the first two elements considered in the selection from group A and B. We denote this global fairness guarantee by δ_{max}. Using ϕ from Theorem 5.1 the cost guarantee is given by $\phi\delta_{max} = \frac{1}{nRel(A)+nRel(B)}\left|\frac{p_A-p_B}{q_A+q_B}\right|\left(\frac{\sigma^{PRP,A}[1]}{nRel(A)} + \frac{\sigma^{PRP,B}[1]}{nRel(B)}\right)$. Further, we show that if the EOR algorithm selects all the elements from one group at some position k, then selecting the remaining elements from the other group satisfies the δ_{max} constraint. We provide a complete proof of this theorem in Appendix C.2. □

We now compare EOR with the Uniform ranking policy and analyze positions k with $\delta = 0$ to avoid discretization effects.

Proposition 5.1 (Costs from EOR vs. Uniform Policy). *The EOR ranking never has higher costs to the groups and total cost to the principal as compared to the Uniform Policy, for those k where $\delta(\sigma_k) = 0$.*

We provide the proof of Proposition 5.1 in Appendix C.3. In summary, we have shown that Algorithm 1 is an efficient algorithm that computes rankings close to the optimal solution, making it a promising candidate for practical use.

6 Extension to G Groups

In this section, we discuss the extension of the EOR algorithm beyond two groups. In particular, we consider the general case where a candidate belongs to one of G groups $g \in [1 \cdots G]$. From Section 4, we can generalize the cost burden to the principal similar to Eq. (4), taking all the groups into account for the normalization factor as follows

$$c(\text{Principal}|\pi_k) = \frac{\sum_i (1 - \mathbb{P}(i \in \sigma_k^{\pi}))\mathbb{P}(r_i = 1|\mathcal{D})}{\sum_{g=1}^{G} nRel(g)}$$

To generalize Algorithm 1 for selecting top k candidates from multiple groups, we define $\delta(\sigma)$ as the EOR criterion that captures the gap between the group with the maximum accumulated relevance proportion and the group with the minimum accumulated relevance proportion,

$$\delta(\sigma) = \max_g \left\{\frac{nRel(g|\sigma)}{nRel(g)}\right\} - \min_g \left\{\frac{nRel(g|\sigma)}{nRel(g)}\right\} \tag{8}$$

The following selection rule then provides the selected group g^* and candidate l_{g^*} to append to the EOR ranking.

$$\begin{aligned} l_g &= \sigma^{PRP,g}[1] \quad \forall g \in \{1 \cdot G\} \\ g^* &= \arg\min_{g \in [1..G]} \delta(\sigma^{EOR} \cup \{l_g\}); \quad l_{g^*} = \sigma^{PRP,g^*}[1] \end{aligned} \tag{9}$$

Note that the above selection rule is a strict generalization of Algorithm 1 and it reflects the intuition of minimizing the gap in relevance proportions for all the groups. It can be verified that the runtime complexity with selection rule according to Eqs. (8), (9) for a constant number of groups G is $O(n \log n + Gn)$. Furthermore, we can extend the cost-approximation guarantee to the multi-group case.

Theorem 6.1 (Global Cost and Fairness Guarantee for multiple groups). *The EOR rankings are cost optimal up to a gap of $\phi\delta(\sigma_k^{EOR})$ for G groups, with $\delta(\sigma_k^{EOR})$ bounded by δ_{max}, such that,*

$$\begin{aligned} \phi &= \frac{2}{(G-1)\sum_{g=1}^{G} nRel(g)} \left(\sum_{\{A,B\}} \left|\frac{p_A - p_B}{q_A + q_B}\right|\right) \forall k \\ \delta_{max} &= \max_g \left\{\frac{\sigma^{PRP,g}[1]}{nRel(g)}\right\} \end{aligned}$$

where $\{A, B\}$ are all G choose 2 possible pairs of groups.

Proof Sketch: We extend the LP formed in Theorem 5.1 to include $G(G-1)$ δ constraints and construct feasible dual variables from the EOR solution for each pair of groups. We then show that the duality gap is bounded by $\phi\delta(\sigma_k^{EOR})$ for a particular prefix k. Note that the ϕ bound for multi-group reduces to the one presented in Theorem 5.1 for two groups. Finally, we present the global a priori bound on $\delta(\sigma_k^{EOR})$ as δ_{max}, which is a strict generalization of the two groups case. We provide complete proof of this theorem in Appendix D.1. □

7 Experimental Evaluation

We now evaluate the EOR framework and algorithm empirically and compare against several baselines – namely Demographic (Statistical) Parity (π^{DP}) [58], FA*IR Ranking Principle (π^{FS}) [61], Probability Ranking Principle (π^{PRP}) [41], Thompson Sampling Policy (π^{TS}) [50], Uniform Policy (π^{unif}), Disparate Treatment of Exposure (π^{EXP}) [49], and Fair Rank Aggregation (π^{RA}) [7] with proportional representation of exposure. We discuss implementation details of these baselines in Appendix E.1.

7.1 Synthetic Data

We first present results on synthetic data where we can control the level of disparate uncertainty. We report a) unfairness and b) effectiveness of rankings for each scenario. The unfairness metric is defined as the area under the curve for the EOR criterion, given by $\sum_{k=1}^{n} |\delta(\sigma_k)|$. To measure the effectiveness of rankings, we report the improvement in total cost over the expected total cost of π^{unif}, computed as $\sum_{k=1}^{n} c(\text{Prinicpal}|\pi_k^{\text{unif}}) - c(\text{Prinicpal}|\pi_k^{(.)})$.

7.1.1 How does π^{EOR} compare against the baselines under varying amounts of disparate uncertainty? Table 1 (left) reports unfairness

	Un-fairness ↓			Effectiveness ↑		
π \Disp. Unc.	High	Medium	Low	High	Medium	Low
π^{EOR}	1.07 ±0.01	1.02 ±0.00	1.02 ±0.00	10.44 ±0.15	11.89 ±0.04	14.58 ±0.10
π^{DP}	11.09 ±0.38	6.02 ±0.07	2.42 ±0.20	10.07 ±0.20	11.33 ±0.04	14.49 ±0.11
π^{PRP}	15.41 ±0.69	7.68 ±0.13	2.63 ±0.17	12.11 ±0.20	12.00 ±0.02	14.62 ±0.09
π^{TS}	11.77 ±0.57	4.96 ±0.07	4.49 ±0.45	7.66 ±0.04	9.62 ±0.06	12.81 ±0.69
π^{unif}	5.96 ±0.13	5.80 ±0.00	6.49 ±0.09	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00
π^{EXP}	9.23 ±0.77	5.62 ±0.01	3.26 ±0.62	11.59 ±0.23	11.97 ±0.03	14.62 ±0.09
π^{RA}	13.97 ±0.71	6.57 ±0.16	2.40 ±0.00	12.02 ±0.19	12.00 ±0.02	14.60 ±0.00
π^{FS}	13.33 ±0.70	7.04 ±0.16	2.95 ±0.17	11.98 ±0.20	12.00 ±0.02	14.62 ±0.09

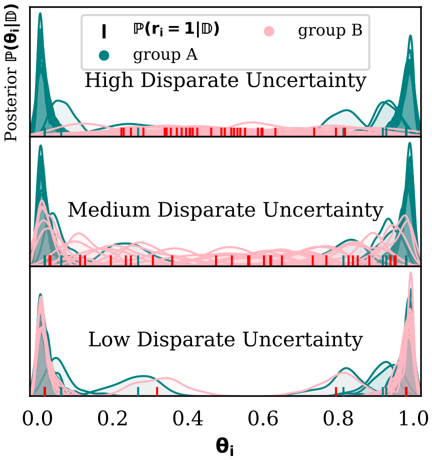


Table 1: Left: Effect of varying disparate uncertainty on Synthetic Dataset, **Right:** Posterior distribution and expected probabilities of relevance shown for a sample from each of high, medium, and low uncertainty setting.

and effectiveness for π^{EOR} and the baselines in terms of mean and standard error over 100 simulations, while Table 1 (right) demonstrates the posterior distribution formed by sampling an instance of each of high, medium and low disparate uncertainty settings. These posterior distributions similar to Figure 3 are for illustrative purposes since only the expected probability of relevance p_i is used for rankings (refer to Section 3.2). The different disparate uncertainty settings are generated synthetically to demonstrate how ranking policies behave if, for example, the Principal collects more data for group B thus reducing the disparate uncertainty among groups. Note, how in the low disparate uncertainty setting, the sharp p_i (close to 0 or 1), would make the identification of relevant candidates easy for both groups. The synthetic generation involves sampling p_i from sharp and flat distributions for group A and B respectively and gradually increasing the sharpness of p_i for group B (implementation details in Appendix E.2).

As predicted by theory, π^{EOR} maintains low unfairness at all levels of disparate uncertainty, outperforming all the baselines π^{PRP}, π^{DP}, π^{TS}, π^{EXP}, π^{RA}, and π^{FS}. Note that π^{EOR} even outperforms the uniform policy π^{unif}, since any individual ranking drawn from π^{unif} is likely to be unfair. In terms of effectiveness, the theoretically optimal skyline is given by π^{PRP}. Across all levels of disparate uncertainty, π^{EOR} is at least competitive with the other baselines, indicating that the EOR fairness does not impose a disproportionate cost of fairness for the Principal.

Note how the gap in the unfairness between π^{EOR} and all other ranking policies is largest when disparate uncertainty is highest. At low levels of disparate uncertainty, π^{EOR} is still more fair as compared to other ranking policies (though the gap in unfairness is smaller) and the effectiveness of π^{EOR} is almost the same as that of π^{PRP}.

7.1.2 At which positions in the rankings do the policies incur unfairness?

While the previous table summarized unfairness across the whole ranking, Figure 4 (left) provides more detailed insights into how unfairness accumulates across positions in the ranking. The only method that is systematically fair across all positions k is π^{EOR}, keeping the unfairness $\delta(\sigma_k)$ from Definition 4.1 close to zero everywhere in the ranking. The baselines generally start accumulating unfairness towards one group right from the top of the ranking. Their unfairness only decreases once they run out of viable candidates from the group they prefer. The only exception is π^{unif}, here for a specific ranking with median $\sum_{k=1}^{n} |\delta(\sigma_k^{\text{unif}})|$. However, rankings from π^{unif} tend to stray much further from zero than the π^{EOR} ranking. Additional results for the medium and low disparate uncertainty settings in Figure 11 of Appendix E.2 further support these findings.

7.1.3 How do the ranking policies distribute the costs between the stakeholders?

In Figure 4 (middle) we investigate how the ranking policies distribute the cost $c(g|\pi_k)$ from Eq. (3) between group A and group B. It shows that only π^{EOR} has an equal cost to both groups across the whole ranking, which can be seen from the overlapping cost curves for both groups. Furthermore, the cost is substantially lower for both groups than their expected cost under the uniform policy (diagonal line).Figure 4 (right) shows the total cost to the principal, and again π^{EOR} is competitive with the baselines.

All other baselines incur substantial disparate costs to the groups, some even worse than the uniform lottery. In particular, π^{DP} selects the candidates alternately between the two groups since group sizes are relatively similar, but this results in selecting a higher proportion of relevance from group A because the relevance probabilities are sharper for group A than for B. As a result, the cost burden is higher for group B. π^{TS} is fairer than π^{PRP}, since it randomizes relevant candidates before sorting them in decreasing order of relevance, however being group oblivious, it still places an uneven cost burden.

The exposure based policies π^{EXP}, π^{RA} motivated by position bias in rankings also do not distribute the costs evenly. π^{EXP} will stochastically allocate most of the top positions to candidates with sharp and high probabilities, close to 1.0 from group A, then to candidates of group B with flat and middle relevance probabilities, and finally the rest of the candidates from group A with sharp but low probabilities, close to 0.0 in the last positions. While this perfectly allocates exposure between group A and B over the full ranking of 61 candidates, group B (the uninformative group) suffers from a high cost burden. Note how the direction of cost burden

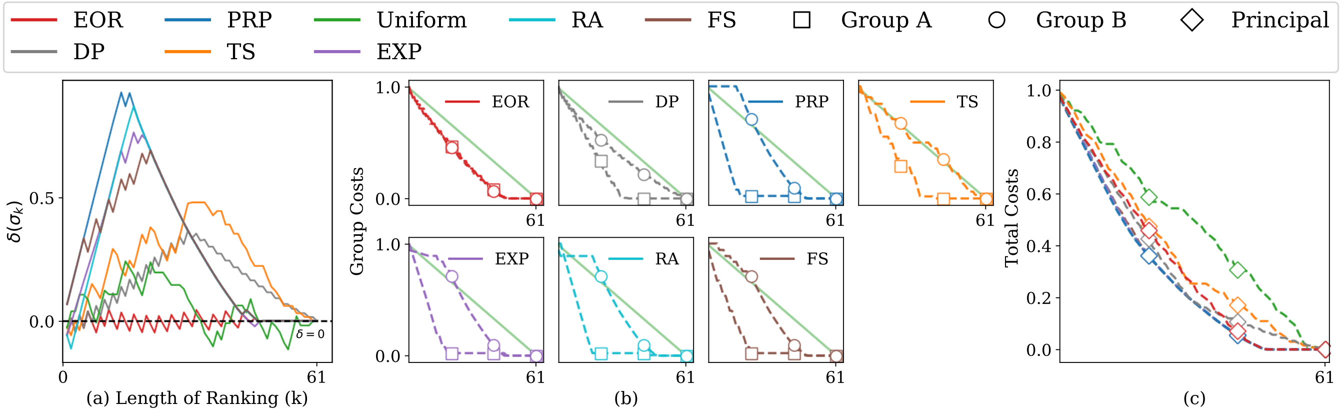


Figure 4: Left: EOR criterion $\delta(\sigma_k)$, **Middle**: group costs according to (3), **Right**: the principal's total cost according to (4) of the ranking policies for the synthetic dataset with high disparate uncertainty shown in top right of Table 1. Group A consists of 30 candidates with sharp probabilities with $p_i \sim \text{Beta}(1/20, 1/20)$. This provides $nRel(A) = 14.96$ expected number of relevant candidates. Group B also has similar candidates, in particular, it has 31 candidates, with relatively flat probabilities $p_i \sim \text{Beta}(5, 5)$, providing $nRel(B) = 14.94$ expected number of relevant candidates.

is opposite to the one π^{EXP} induced in the example of Figure 1, where group B was smaller in size to group A.

7.2 US Census Survey Data

While the synthetic experiments provide insights into the behavior of ranking policies under varying conditions, we now investigate how far π^{EOR} can mitigate unfairness as it arises in real-world datasets where the relevance probabilities $\mathbb{P}(r_i|\mathcal{D})$ are learned from data. In particular, we consider the US Census Survey dataset [14] for the year 2018 and the state of Alabama and New York, consisting of 22,268 and 103,021 records respectively. The task is to predict whether the income for an individual $> \$50K$ based on features such as educational attainment, occupation, class of worker etc. We use this task as a stand-in for some task where individuals receive a benefit from being evaluated positively. To get group-calibrated estimates of $\mathbb{P}(r_i|\mathcal{D})$, we train a gradient boosting classifier followed by Platt Scaling on the validation subset of the data. We evaluate the EOR criterion and costs on the test subset of these records. Full details for dataset pre-processing and training can be found in Appendix E.3. Because these rankings are large (up to $\sim 20K$ size), π^{EXP} and π^{FS} are not computationally tractable. π^{RA} performs similarly to π^{PRP} and we include it in Appendix E.3 for completeness.

7.2.1 How do the ranking policies compare when using learned probability estimates? To evaluate the two-group EOR algorithm, we first only rank individuals labeled as White and Black or African American. Figure 5 (top) shows that EOR ranking is effective even with estimated probabilities. In particular, while the ranking algorithms only use estimated probabilities, the EOR criterion, and costs are evaluated on the true relevance labels from the test set. Nevertheless, π^{EOR} still evaluates δ close to zero and distributes costs among the stakeholders more evenly than the other baseline policies π^{PRP}, π^{DP}, and even π^{unif}, π^{TS} for a specific ranking with median $\sum_{k=1}^{n} |\delta(\sigma_k)|$. Additional experiments in Appendix E.3 further confirm these findings.

7.2.2 How does EOR Ranking perform for more than two groups? Figure 5 (bottom) shows results on the US Census Dataset for four groups, again using estimated relevances for ranking but evaluating against the true relevance labels from the test dataset. Note that for more than two groups, the EOR constraint defined according to (8) will always be non-negative as it measures the absolute difference in relevance proportions between the groups that are furthest apart. We observe that similar to the results with two groups, the EOR ranking keeps the unfairness δ lower (close to zero) as compared to other policies in Figure 5 (left). Additionally, π^{EOR} also distributes the costs evenly among all stakeholders for the generalized case of more than two groups, as noted by the overlapping of dashed lines for the four group costs (middle). Finally, π^{EOR} is competitive with the optimal π^{PRP} in terms of total cost for the principal.

7.3 Amazon Shopping Audit

In the final experiment, we investigate how the EOR framework can be used for auditing. To illustrate this point, we use a dataset of Amazon shopping queries [39], which includes a baseline model for predicting the relevance of products given a search query. We further augment this dataset with logged rankings from the Amazon website as collected for the Markup report [60], which investigated Amazon's placement of its own brand products as compared to other brands based on star ratings, reviews etc. The Markup data consists of popular search query-product pairs along with logged rankings of these products on Amazon's platform, but it does not contain human-annotated relevance labels. We focus the audit on bias between the group of Amazon-owned brands (group A) or any other brand (group B). As the first step of the audit, we calibrate p_i by fitting a Platt-scaling calibrator using validation data for both groups. Figure 6a shows that the calibrated p_i on the test dataset binned across 20 equal-sized bins, lies close to the perfectly calibrated line. As the second step of the audit, we use the Markup dataset with logged rankings[4] and compute p_i using the calibrated

[4] https://github.com/the-markup/investigation-amazon-brands

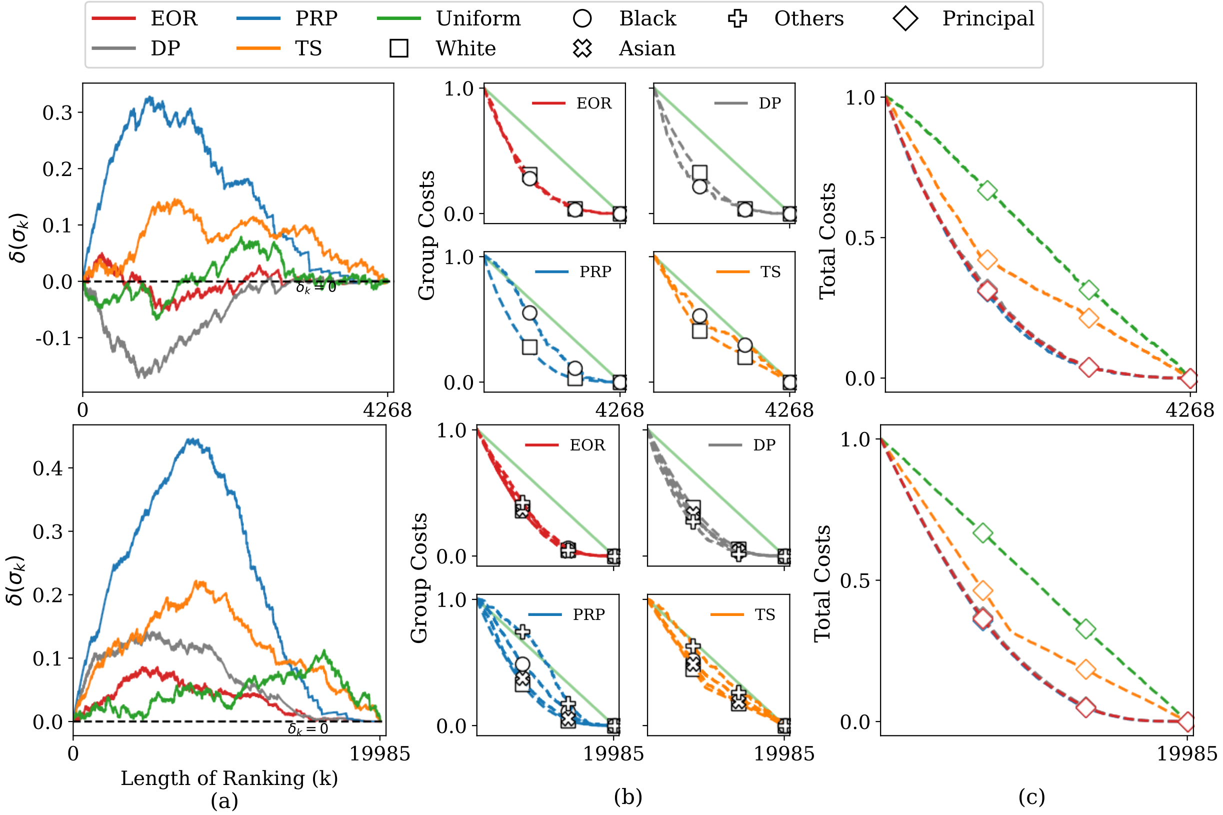


Figure 5: US Census Dataset: EOR criterion $\delta(\sigma_k)$ and cost of the ranking policies computed with true relevance labels from the test subset for the US Census dataset. **Top: Two groups setting** using the White and Black/African American racial groups for the state of Alabama. **Bottom: Multiple (four groups) setting** using White, Black/African American, Asian, and Other for the state of NY.

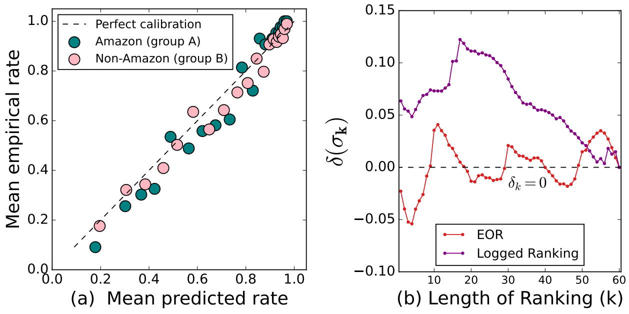


Figure 6: Left: Group-wise calibration of $\mathbb{P}(r_i|\mathcal{D})$ for Amazon shopping queries on the test set according to the baseline model after Platt Scaling. **Right:** Fairness of logged Amazon rankings compared to EOR rankings in terms of $\delta(\sigma_k)$ averaged over queries.

baseline relevance prediction model. The EOR criterion (7) is averaged over queries for the logged rankings, and the EOR rankings are produced by Algorithm 1. Figure 6b shows that there exists a ranking σ^{EOR} that has $\delta(\sigma_k^{EOR})$ closer to zero for most prefix k. The logged rankings from Amazon's platform show estimated $\delta(\sigma_k)$ that are farther away from zero for at least some prefixes of k, reflecting a potential favoring of Amazon brand products. A limitation of this analysis is that unlike in a real audit where the auditor has access to the production model of p_i, our baseline model may be subject to hidden confounding, and thus does not provide conclusive evidence of unfairness. In particular, the production rankings may depend on other features beyond product titles (e.g, product descriptions, bullet points, star ratings, etc.). However, the analysis does demonstrate how the EOR criterion can be used for auditing, if the auditor is given access to the production ranking model to avoid confounding. We provide further details in Figure E.4 and our source code with experiment implementation can be found here.[5]

8 Conclusion

This paper studies the problem of disparate uncertainty across groups as a source of unfairness in ranking when these rankings are used as part of a human decision-making process. In particular, this paper introduces a framework that formalizes this unfairness by relating it both to a fair lottery and to the costs that an imperfect model imposes on the various stakeholders. Recognizing that it may be difficult to avoid disparate uncertainty in real-world models, the paper develops the EOR procedure to produce rankings that provably mitigate the effects of disparate uncertainty between groups. Beyond its strong theoretical guarantees, we find that the

[5] https://github.com/RichRast/DisparateUncertainty

EOR method outperforms existing methods for fair ranking across a wide range of settings. Furthermore, we illustrate that the EOR criterion can also be used as a tool to audit a real-world system. We conjecture that this combination of theoretical grounding, computational efficiency, and strong empirical performance provides viable conditions for making the proposed framework and algorithm accessible for thoughtful use in practice.

9 Ethical Considerations

This work explicitly addresses the potentially negative societal impact of machine learning predictions that include disparities between groups in the context of ranking interfaces. However, as pointed out by previous research [34, 46], we do not prescribe distilling down the fairness of a system into a single metric – the fairness criterion we propose. We emphasize that it is important to carefully consider the domain specifics and the particular situation where our method may be deployed.

We also note that while our EOR algorithm does not worsen the fairness within each group (i.e., within group ordering is maintained), it doesn't improve within-group fairness either. Exploring this dichotomy of satisfying within and between group fairness simultaneously in the presence of differential uncertainty is an important open question.

Acknowledgments

This research was supported in part by NSF Awards IIS-2008139 and IIS-2312865. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors. We thank Kate Donahue, Marios Papachristou, Aaron Tucker, Sarah Dean, Luke Wang, Emily Ryu, Ashudeep Singh, Taran Pal Singh, and Woojeong Kim for helpful comments and discussions. We also thank the anonymous reviewers at the Epistemic AI, UAI workshop for helpful feedback.

References

[1] Kenneth Arrow. 1971. *The Theory of Discrimination.* Working Papers 403. Princeton University, Department of Economics, Industrial Relations Section. https://EconPapers.repec.org/RePEc:pri:indrel:30a

[2] Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. 2020. Equalized odds postprocessing under imperfect group information. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 108)*, Silvia Chiappa and Roberto Calandra (Eds.). PMLR, 1770–1780. https://proceedings.mlr.press/v108/awasthi20a.html

[3] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* 50, 1 (2021), 3–44. https://doi.org/10.1177/0049124118782533 arXiv:https://doi.org/10.1177/0049124118782533

[4] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. *CoRR* abs/1805.01788 (2018). arXiv:1805.01788 http://arxiv.org/abs/1805.01788

[5] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[6] Robin Burke. 2017. Multisided Fairness for Recommendation. *CoRR* abs/1707.00093 (2017). arXiv:1707.00093 http://arxiv.org/abs/1707.00093

[7] Kathleen Cachel and Elke Rundensteiner. 2023. Fairer Together: Mitigating Disparate Exposure in Kemeny Rank Aggregation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1347–1357. https://doi.org/10.1145/3593013.3594085

[8] L. Elisa Celis, Chris Hays, Anay Mehrotra, and Nisheeth K. Vishnoi. 2021. The Effect of the Rooney Rule on Implicit Bias in the Long Term. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 678–689. https://doi.org/10.1145/3442188.3445930

[9] L. Elisa Celis, Anay Mehrotra, and Nisheeth K. Vishnoi. 2020. Interventions for Ranking in the Presence of Implicit Bias. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 369–380. https://doi.org/10.1145/3351095.3372858

[10] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. 2017. Ranking with Fairness Constraints. In *International Colloquium on Automata, Languages and Programming*.

[11] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. https://doi.org/10.1089/BIG.2016.0047

[12] Brian Collins. 2007. Tackling Unconscious Bias in Hiring Practices: The Plight of the Rooney Rule. *NYU Law Review* 82 (06 2007).

[13] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) *(KDD '17)*. Association for Computing Machinery, New York, NY, USA, 797–806. https://doi.org/10.1145/3097983.3098095

[14] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 6478–6490. https://proceedings.neurips.cc/paper_files/paper/2021/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf

[15] Ezekiel Dixon-Roman, Howard Everson, and John Mcardle. 2013. Race, Poverty and SAT Scores: Modeling the Influences of Family Income on Black and White High School Students' SAT Performance. *Teachers College Record* 115 (05 2013). https://doi.org/10.1177/016146811311500406

[16] Marina Drosou, H.V. Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. 2017. Diversity in Big Data: A Review. *Big Data* 5, 2 (2017), 73–84. https://doi.org/10.1089/big.2016.0054 arXiv:https://doi.org/10.1089/big.2016.0054 PMID: 28632443.

[17] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Cambridge, Massachusetts) *(ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[18] Vitalii Emelianov, Nicolas Gast, Krishna P. Gummadi, and Patrick Loiseau. 2020. On Fair Selection in the Presence of Implicit Variance. In *Proceedings of the 21st ACM Conference on Economics and Computation* (Virtual Event, Hungary) *(EC '20)*. Association for Computing Machinery, New York, NY, USA, 649–675. https://doi.org/10.1145/3391403.3399482

[19] Vitalii Emelianov, Nicolas Gast, Krishna P. Gummadi, and Patrick Loiseau. 2022. On fair selection in the presence of implicit and differential variance. *Artificial Intelligence* 302 (2022), 103609. https://doi.org/10.1016/j.artint.2021.103609

[20] Anthony W. Flores, Kristin Bechtel, and Christopher T. Lowenkamp. 2016. False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used across the Country to Predict Future Criminals. and It's Biased against Blacks". *Federal Probation* 80 (2016), 38. https://www.uscourts.gov/federal-probation-journal/2016/09/false-positives-false-negatives-and-false-analyses-rejoinder

[21] Nikhil Garg, Hannah Li, and Faidra Monachou. 2021. Standardized Tests and Affirmative Action: The Role of Bias and Variance. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 261. https://doi.org/10.1145/3442188.3445889

[22] Barbara Goodwin. 1992. *Justice by Lottery.* University of Chicago Press, Chicago.

[23] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

[24] Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness Without Demographics in Repeated Loss Minimization. In *International Conference on Machine Learning*.

[25] Li He and Toni M. Whited. 2023. Underrepresentation of Women CEOs. (2023). Available at SSRN: https://ssrn.com/abstract=4615373 or http://dx.doi.org/10.2139/ssrn.4615373.

[26] Tim Henning. 2015. From Choice to Chance? Saving People, Fairness, and Lotteries. *Philosophical Review* 124, 2 (2015), 169–206. https://doi.org/10.1215/00318108-2842176

[27] Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning* 110, 3 (2021), 457–506. https://doi.org/10.1007/s10994-021-05946-3

[28] Tim De Jonge and Djoerd Hiemstra. 2023. UNFair: Search Engine Manipulation, Undetectable by Amortized Inequity *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA.

[29] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 67)*, Christos H. Papadimitriou (Ed.). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 43:1–43:23. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[30] Jon Kleinberg and Manish Raghavan. 2018. Selection Problems in the Presence of Implicit Bias. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[31] Nikola Konstantinov and Christoph H. Lampert. 2021. Fairness Through Regularization for Learning to Rank. *CoRR* abs/2102.05996 (2021). arXiv:2102.05996 https://arxiv.org/abs/2102.05996

[32] Michele Loi and Christoph Heitz. 2022. Is calibration a fairness requirement? An argument from the point of view of moral philosophy and decision theory. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (<conf-loc>, <city>Seoul</city>, <country>Republic of Korea</country>, </conf-loc>) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2026–2034. https://doi.org/10.1145/3531146.3533245

[33] Anay Mehrotra and Nisheeth K Vishnoi. 2022. Fair Ranking with Noisy Protected Attributes. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=mTra5BIUyRV

[34] Jakob Mökander, Jessica Morley, Mariarosaria Taddeo, and L. Floridi. 2021. Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. *Science and Engineering Ethics* 27 (2021).

[35] Harikrishna Narasimhan, Andy Cotter, Maya Gupta, and Serena Lutong Wang. 2020. Pairwise Fairness for Ranking and Regression. In *33rd AAAI Conference on Artificial Intelligence*.

[36] Edmund S. Phelps. 1972. The Statistical Theory of Racism and Sexism. *The American Economic Review* 62, 4 (1972), 659–661. http://www.jstor.org/stable/1806107

[37] J. Platt. 2000. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In *Advances in Large Margin Classifiers*.

[38] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On Fairness and Calibration. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/b8b9c74ac526fffbeb2d39ab038d1cd7-Paper.pdf

[39] Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. 2022. Shopping Queries Dataset: A Large-Scale ESCI Benchmark for Improving Product Search. (2022). arXiv:2206.06588

[40] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[41] Stephen Robertson. 1977. The Probability Ranking Principle in IR. *Journal of Documentation* 33 (12 1977), 294–304. https://doi.org/10.1108/eb026647

[42] Yuta Saito and Thorsten Joachims. 2022. Fair Ranking as Fair Division: Impact-Based Individual Fairness in Ranking. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Washington DC, USA) *(KDD '22)*. Association for Computing Machinery, New York, NY, USA, 1514–1524. https://doi.org/10.1145/3534678.3539353

[43] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the Impact of User Attentionon Fair Group Representation in Ranked Lists. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 553–562. https://doi.org/10.1145/3308560.3317595

[44] Ben Saunders. 2008. The Equality of Lotteries. *Philosophy* 83, 3 (2008), 359–372. https://doi.org/10.1017/s0031819108000727

[45] Ben Saunders. 2018. Equality in the Allocation of Scarce Vaccines. *Les Ateliers de l'Éthique / the Ethics Forum* 13, 3 (2018), 65–84. https://doi.org/10.7202/1061219ar

[46] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[47] Zeyu Shen, Zhiyi Wang, Xingyu Zhu, Brandon Fain, and Kamesh Munagala. 2023. Fairness in the Assignment Problem with Uncertain Priorities. arXiv:2301.13804 [cs.GT]

[48] Ashudeep Singh and Thorsten Joachims. 2017. Equality of Opportunity in Rankings.

[49] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2219–2228. https://doi.org/10.1145/3219819.3220088

[50] Ashudeep Singh, David Kempe, and Thorsten Joachims. 2021. Fairness in Ranking under Uncertainty. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., 11896–11908. https://proceedings.neurips.cc/paper_files/paper/2021/file/63c3ddcc7b23daa1e42dc41f9a44a873-Paper.pdf

[51] Rachael Tatman. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain, 53–59. https://doi.org/10.18653/v1/W17-1606

[52] Lequn Wang and Thorsten Joachims. 2021. User Fairness, Item Fairness, and Diversity for Rankings in Two-Sided Markets. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval* (Virtual Event, Canada) *(ICTIR '21)*. Association for Computing Machinery, New York, NY, USA, 23–41. https://doi.org/10.1145/3471158.3472260

[53] Lequn Wang and Thorsten Joachims. 2023. Uncertainty Quantification for Fairness in Two-Stage Recommender Systems. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining* (Singapore, Singapore) *(WSDM '23)*. Association for Computing Machinery, New York, NY, USA, 940–948. https://doi.org/10.1145/3539597.3570469

[54] Lequn Wang, Thorsten Joachims, and Manuel Gomez Rodriguez. 2022. Improving Screening Processes via Calibrated Subset Selection. In *International Conference on Machine Learning*.

[55] Dong Wei, Md Mouinul Islam, Baruch Schieber, and Senjuti Basu Roy. 2022. Rank Aggregation with Proportionate Fairness. In *Proceedings of the 2022 International Conference on Management of Data* (Philadelphia, PA, USA) *(SIGMOD '22)*. Association for Computing Machinery, New York, NY, USA, 262–275. https://doi.org/10.1145/3514221.3517865

[56] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. 2019. Predictive Inequity in Object Detection. arXiv:1902.11097 [cs.CV]

[57] Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. 2019. Balanced Ranking with Diversity Constraints. 6035–6042. https://doi.org/10.24963/ijcai.2019/836

[58] Ke Yang and Julia Stoyanovich. 2016. Measuring Fairness in Ranked Outputs. *CoRR* abs/1610.08559 (2016). arXiv:1610.08559 http://arxiv.org/abs/1610.08559

[59] Tao Yang, Zhichao Xu, Zhenduo Wang, Anh Tran, and Qingyao Ai. 2023. Marginal-Certainty-Aware Fair Ranking Algorithm. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining* (Singapore, Singapore) *(WSDM '23)*. Association for Computing Machinery, New York, NY, USA, 24–32. https://doi.org/10.1145/3539597.3570474

[60] Leon Yin and Adrianne Jeffries. 2021. How We Analyzed Amazons Treatment of its Brands in Search Results. *The Markup* (10 2021). https://tinyurl.com/markup-amazon

[61] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Singapore, Singapore) *(CIKM '17)*. Association for Computing Machinery, New York, NY, USA, 1569–1578. https://doi.org/10.1145/3132847.3132938

[62] Meike Zehlike and Carlos Castillo. 2020. Reducing Disparate Exposure in Ranking: A Learning To Rank Approach. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 2849–2855. https://doi.org/10.1145/3366424.3380048

[63] Meike Zehlike, Tom Sühr, Ricardo Baeza-Yates, Francesco Bonchi, Carlos Castillo, and Sara Hajian. 2022. Fair Top-k Ranking with Multiple Protected Groups. *Inf. Process. Manage.* 59, 1 (jan 2022), 28 pages. https://doi.org/10.1016/j.ipm.2021.102707

[64] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2021. Fairness in Ranking: A Survey. *CoRR* abs/2103.14000 (2021). arXiv:2103.14000 https://arxiv.org/abs/2103.14000

A Notation Summary

n	number of candidates
$i \in \{1, \cdots, n\}$	candidate
G	number of groups
$g \in \{1, 2, \cdots G\}$	group
k	ranking prefix
$S(g)$	size of group g
$nRel(g) \in \mathbb{R}$	expected number of relevant candidates for group g

$r_i \in \{0, 1\}$	binary relevance of candidate i
$\theta_i \in [0, 1]$	probability of relevance of candidate i
$\mathcal{D}$	historical data
$\mathbb{P}(\theta_i \mid \mathcal{D})$	posterior distribution
$p_i = \mathbb{P}(r_i \mid \mathcal{D}) \in [0, 1]$	expected probability of relevance of candidate i
$P = (p_i)_{i \in \{1, \cdots, n\}}$	relevance probability vector
X	vector indicating whether candidate i was selected
$\mathbb{1}_g \in \{0, 1\}^n$	indicator if candidate i belongs to group g

π	policy
σ_k^π	top k ranking $\sigma_k \sim \pi$
$\sigma^{PRP,g}[i]$	i^{th} candidate in the PRP ranking of group g
$\delta(\sigma)$	EOR measure for ranking σ

B Extended Related Work

Our work complements and extends prior research on fairness in rankings [64]. The classical fairness desiderata considered are variations of *proportional representation* [17, 58]. Broadly, proportional representation ensures representation by group size in top k selection or at every prefix k of the ranking. Other popular notions include diversity based constraints [10, 16, 57] like Rooney Rule and affirmative action that ensure representation of the designated disadvantaged group, and threshold based formulations [54, 61, 63] that ensure a minimum number of candidates to be selected from the disadvantaged group.

Another prominent class of fairness notions in rankings corresponds to *exposure* based formulations. Exposure [43, 49, 62] quantifies the amount of attention allocated to candidates individually or from a particular group. These formulations include equity of exposure, disparate treatment of exposure that allocates exposure proportional to amortized relevance, and disparate impact of exposure that allocates exposure proportional to impact (e.g. economic impact of ranking) among other variations. See [4] for a similar concept of equity of attention. *Proportional representation*, *diversity constraints*, and *exposure* are motivated by representation by group size, normative designation of disadvantaged group, and allocation of attention respectively. Our work, on the other hand, is motivated by unfairness due to differential uncertainty between groups and is grounded in the axiomatic fairness of a lottery system.

Our problem setup involves aggregating candidates from groups and while research on fair rank aggregation appears related, the goal there is much different. In particular, fair rank aggregation achieves maximum consensus accuracy when multiple voters rank all candidates subject to fairness constraints of group exposure [7] or p-fairness [55]. Work on multi sided fairness [6, 52] similarly considers diversity constraints or exposure-based formulations. Finally, while [4, 31, 42, 50] propose an amortized notion of fairness, our work proposes a non-amortized fairness criterion at every position k of the ranking.

Recently, there has been a growing interest in the study of fairness in rankings under uncertainty. The classical desideratum in this literature studies the relation of group-wise calibration for fairness [11, 20, 29, 32, 38]. Our work is orthogonal to this discussion. In particular, we only assume that calibrated probability of relevance is given and instead focus on how differential sharpness of probabilities can cause unfairness. [50] introduced an approximate notion of fairness that is violated if the principal ranks candidates that appear more than a certain proportion of their estimated relevance distribution. One way to achieve this in expectation is through randomization of relevances drawn from the predictive posterior distribution. Other works have introduced methods that quantify uncertainty in rankings [53] to update and learn better estimates of relevances iteratively [59]. These works do not consider the unfairness caused due to differential uncertainty between groups. While methods that reduce uncertainty for all groups are needed, we also need to account for unfairness due to the existing disparate uncertainty that is unfortunately widespread in practical settings.

Another line of research focuses on statistical discrimination and the study of noisy estimates of relevances for selection problems [1, 36]. This literature establishes that the differential accuracy of models causes unfairness [5, 15, 24, 51, 56] for individuals based on their group membership. Recently, [19, 21] studied the role of affirmative action in the presence of differential variance between groups in rankings. Their method [19] corrects the bias in noisy relevance estimates given the variance of the true relevance distribution. Fairness in selection processes has also been extensively studied in the presence of group-based implicit bias [8, 9, 18, 30], uncertainty in preferences [47] and in the presence of noisy sensitive attributes [33]. This line of research analyzes the effect of affirmative actions like the Rooney rule on the utility to the principal or how implicit bias affects the diversity of the selection set.

Our work is also motivated by Equality of opportunity framework, first introduced by [23] in the classification setting. It has provided a compelling notion of balancing the cost burden among stakeholders [2, 11, 13]. For rankings, there has been some work in transferring the idea of equalized odds with learning a ranking function during training [62] to reduce disparate exposure or augmenting the training loss with regularizers that minimize costs for both groups [31, 35]. Our work extends this literature to introduce a framework connecting the unfairness in rankings due to the disparate uncertainty to the distribution of cost burden among stakeholders by anchoring on the fairness of random lottery.

C Proofs

C.1 Proof of Theorem 5.1

Proof. We use linear duality for proving this theorem. In order to find a lower bound on the cost optimal ranking that satisfies the EOR fairness constraint, we relax the corresponding Integer Linear Problem (ILP) to a Linear Program (LP) by turning any integer constraints $X \in \{0, 1\}$ in the primal into $0 \leq X \leq 1$. For the relaxed LP, we formulate its dual and construct a set of dual variables λ corresponding to the solution from the EOR Algorithm. With the dual solution of EOR and the relaxed LP solution, we obtain an upper bound of the duality gap. Since the upper bound on this duality gap is w.r.t. the relaxed LP, it will also be an upper bound for the optimal ILP.

We define the primal of the LP for finding a solution X as follows

$$\max_{X \geq 0} \quad f(X) = \frac{P^T X}{nRel(A) + nRel(B)} \tag{Primal}$$

$$\text{s.t.} \quad X \leq 1 \tag{10}$$

$$X^T \mathbb{1} \leq k \tag*{(select up to k elements)}$$

$$Q_{A,B}^T X \leq \delta(\sigma_k^{EOR}) \tag{11}$$

$$Q_{B,A}^T X \leq \delta(\sigma_k^{EOR}) \tag{12}$$

We define $Q_{A,B} \in \mathbb{R}^n$ where each element of $Q_{A,B}$ is $q_i(\mathbb{1}_A - \mathbb{1}_B)_i$, $q_{i \in g} = \frac{p_i}{nRel(g)}$ and $Q_{A,B} = -Q_{B,A}$. Note that the Primal objective is equivalent to minimizing the total cost = $1 - \frac{P^T X}{nRel(A)+nRel(B)}$.

The first constraint (10) ensures valid values for X (with corresponding dual variables λ_i'). The second constraint is for selecting k candidates (dual variable λ_k) and the last two constraints (11) and (12) ensure that the ranking solution is EOR-fair optimal (dual variables $\lambda_{A,B}$, $\lambda_{B,A}$). The Dual LP is formed as follows

$$\min_{\lambda \geq 0} \quad g(\lambda) = \delta(\sigma_k^{EOR})(\lambda_{A,B} + \lambda_{B,A}) + k\lambda_k + \sum_{i=1}^{n} \lambda_i' \tag{Dual}$$

$$\text{s.t.} \quad Q_{A,B}(\lambda_{A,B} - \lambda_{B,A}) + \lambda_k + \lambda' \geq \frac{P}{nRel(A) + nRel(B)} \tag{13}$$

We construct a feasible point of the dual from the EOR solution as follows. The key insight here is to reason w.r.t the last elements selected (or the first elements available if no element from the group has been selected) by the EOR Algorithm at prefix k from each of the groups A and B, namely k_A, k_B respectively.

$$\lambda_{A,B} = \frac{1}{nRel(A) + nRel(B)} \left[\frac{p_A - p_B}{q_A + q_B} \right]_+ \tag{14}$$

$$\lambda_{B,A} = \frac{1}{nRel(A) + nRel(B)} \left[- \left(\frac{p_A - p_B}{q_A + q_B} \right) \right]_+ \tag{15}$$

Using (14) and (15) we know that only ever one of $\lambda_{A,B}$ or $\lambda_{B,A}$ is non zero. If $p_A \geq p_B$, then $\lambda_{A,B} \geq 0$ and $\lambda_{B,A} = 0$. Similarly, if $p_B \geq p_A$, then $\lambda_{B,A} \geq 0$ and $\lambda_{A,B} = 0$.

We construct λ_k and λ'_i as follows

$$\lambda_k \;=\; \left[\frac{p_A}{nRel(A)+nRel(B)} - q_A(\lambda_{A,B}-\lambda_{B,A})\right] = \left[\frac{p_B}{nRel(A)+nRel(B)} - q_B(\lambda_{B,A}-\lambda_{A,B})\right] \tag{16}$$

$$\lambda'_{i\in A} \;=\; \left[\frac{p_i}{nRel(A)+nRel(B)} - \lambda_k - q_i(\lambda_{A,B}-\lambda_{B,A})\right]_+ \tag{17}$$

$$\lambda'_{i\in B} \;=\; \left[\frac{p_i}{nRel(A)+nRel(B)} - \lambda_k - q_i(\lambda_{B,A}-\lambda_{A,B})\right]_+ \tag{18}$$

We prove that the constructed dual variables λ are non-negative in Lemma 5.2 and that $\lambda' = 0$ for any element not selected in the EOR ranking. In Lemma 5.3, we prove that the constructed dual variables λ are feasible. Given the feasibility of dual variables, we analyze the **duality gap** given by

$$g(\lambda^*) - f(X) = \delta(\sigma_k^{EOR})(\lambda_{A,B}+\lambda_{B,A}) + k\lambda_k + \sum_{i=1}^{n}\lambda'_i - \frac{P^TX}{nRel(A)+nRel(B)}$$

From Lemma 5.2, $\lambda'_i = 0$ for $i > k_A$, $\lambda'_j = 0$ for $j > k_B$, where k_A elements are selected from group A, k_B from group B by the EOR Algorithm and $k = k_A + k_B$. Substituting the values for λ' from (17), (18), the duality gap is

$$\begin{aligned} &= \delta(\sigma_k^{EOR})(\lambda_{A,B}+\lambda_{B,A}) + k\lambda_k + \sum_{i=1}^{k_A}\left(\frac{p_i}{nRel(A)+nRel(B)} - \lambda_k - q_i(\lambda_{A,B}-\lambda_{B,A})\right) \\ &+ \sum_{j=1}^{k_B}\left(\frac{p_j}{nRel(A)+nRel(B)} - \lambda_k - q_j(\lambda_{B,A}-\lambda_{A,B})\right) - \frac{P^TX}{nRel(A)+nRel(B)} \end{aligned}$$

We know that $\sum_{i=1}^{k_A}\lambda_k + \sum_{j=1}^{k_B}\lambda_k = k\lambda_k$ and $P^TX = \sum_{i=1}^{k_A}p_i + \sum_{j=1}^{k_B}p_j$. Further, only one of $\lambda_{A,B}$ or $\lambda_{B,A}$ is non-negative according to (14), (15).

If $\lambda_{A,B} \geq 0$, then the duality gap can be written as

$$= \delta(\sigma_k^{EOR})\lambda_{A,B} - \sum_{i=1}^{k_A}q_i\lambda_{A,B} + \sum_{j=1}^{k_B}q_j\lambda_{A,B} = \lambda_{A,B}\left(\delta(\sigma_k^{EOR}) - \left(\sum_{i=1}^{k_A}q_i - \sum_{j=1}^{k_B}q_j\right)\right)$$

Since we have $-\delta(\sigma_k^{EOR}) \leq \sum_{i=1}^{k_A}q_i - \sum_{j=1}^{k_B}q_j \leq \delta(\sigma_k^{EOR})$ from Lemma 5.1,

$$\text{Duality Gap} = \lambda_{A,B}\left(\delta(\sigma_k^{EOR}) - \left(\sum_{i=1}^{k_A}q_i - \sum_{j=1}^{k_B}q_j\right)\right) \leq 2\lambda_{A,B}\delta(\sigma_k^{EOR}) \tag{19}$$

If $\lambda_{B,A} \geq 0$, then the duality gap can be written as

$$= \delta(\sigma_k^{EOR})\lambda_{B,A} + \sum_{i=1}^{k_A}q_i\lambda_{B,A} - \sum_{j=1}^{k_B}q_j\lambda_{B,A} = \lambda_{B,A}\left(\delta(\sigma_k^{EOR}) + (\sum_{i=1}^{k_A}q_i - \sum_{j=1}^{k_B}q_j)\right)$$

and again, since $-\delta(\sigma_k^{EOR}) \leq \sum_{i=1}^{k_A}q_i - \sum_{j=1}^{k_B}q_j \leq \delta(\sigma_k^{EOR})$ from Lemma 5.1,

$$\text{Duality Gap} = \lambda_{B,A}\left(\delta(\sigma_k^{EOR}) + (\sum_{i=1}^{k_A}q_i - \sum_{j=1}^{k_B}q_j)\right) \leq 2\lambda_{B,A}\delta(\sigma_k^{EOR}) \tag{20}$$

From Eqs. (14), (15), (19), (20) , the duality gap between EOR solution and the optimal solution is bounded by

$$\frac{2\delta(\sigma_k^{EOR})}{nRel(A)+nRel(B)}\left|\frac{p_A-p_B}{q_A+q_B}\right|$$

This proves that the EOR solution can only be ever as worse as $\phi\delta(\sigma_k^{EOR})$ when compared with the optimal solution, where $\phi = \frac{2}{nRel(A)+nRel(B)}\left|\frac{p_A-p_B}{q_A+q_B}\right|$ □

Lemma 5.1. *EOR ranking is $\delta(\sigma_k^{EOR})$ fairness optimal, implying that $-\delta(\sigma_k^{EOR}) \leq \sum_{i=1}^{k_A}q_i - \sum_{j=1}^{k_B}q_j \leq \delta(\sigma_k^{EOR})$.*

Since $\sum_{i=1}^{k_A}q_i - \sum_{j=1}^{k_B}q_j = \frac{nRel(A|\sigma_k)}{nRel(A)} - \frac{nRel(B|\sigma_k)}{nRel(B)}$, the lemma follows directly from the definition of $\delta(\sigma_k)$ in Eq. (7) and the EOR ranking principle of choosing the candidate that minimizes $\delta(\sigma_k)$.

□

Lemma 5.2. The constructed dual variables $\lambda \geq 0$. In particular, for any $i > k_A$ in group A and $j > k_B$ in group B, it holds that $\lambda'_i = 0$ and $\lambda'_j = 0$ and for any $i \leq k_A$ and $j \leq k_B$, it holds that $\lambda'_i \geq 0$ and $\lambda'_j \geq 0$.

Proof. In this Lemma, we show that $\lambda' = 0$ for the elements not selected by the EOR Algorithm and $\lambda' \geq 0$ for the elements that were selected. Without loss of generality, we consider the element at index i that belongs to group A.

$$\begin{aligned}
\lambda'_{i \in A} &= \left[\frac{p_i}{nRel(A) + nRel(B)} - \lambda_k - q_i(\lambda_{A,B} - \lambda_{B,A})\right]_+ \\
&= \left[\frac{p_i}{nRel(A) + nRel(B)} - \frac{p_A}{nRel(A) + nRel(B)} + q_A(\lambda_{A,B} - \lambda_{B,A}) - q_i(\lambda_{A,B} - \lambda_{B,A})\right]_+ \\
&= \left[\frac{p_i - p_A}{nRel(A) + nRel(B)} + (q_i - q_A)(\lambda_{B,A} - \lambda_{A,B})\right]_+ \qquad (21)
\end{aligned}$$

The second equality above is obtained by substituting λ_k from Eq. (16) and the last equality by rearranging. We now consider two cases – for elements not selected and selected by the EOR Algorithm respectively.

Case I: Elements not selected by the EOR Algorithm.

We have i) $p_i \leq p_A$ and $q_i \leq q_A$ as EOR selects in decreasing order of probabilities, and ii) either $\lambda_{A,B} \geq 0$ or $\lambda_{B,A} \geq 0$ as only one of them can be nonzero from (14), (15).

In Eq. (21), if $\lambda_{B,A} \geq 0$, then $\lambda_{A,B} = 0$ and with $p_i \leq p_A$, $q_i \leq q_A$ the resultant quantity would be negative, which would result in λ'_i clipped to 0.

$$\begin{aligned}
\lambda'_i &= \left[\frac{p_i - p_A}{nRel(A) + nRel(B)} + (q_i - q_A)\lambda_{B,A}\right]_+ \\
&\leq 0
\end{aligned}$$

In Eq. (21), if $\lambda_{A,B} \geq 0$, then $\lambda_{B,A} = 0$. We can then substitute $\lambda_{A,B} = \frac{1}{nRel(A)+nRel(B)}\left(\frac{p_A - p_B}{q_A + q_B}\right)$ in Eq. (21),

$$\begin{aligned}
\lambda'_i &= \left[\frac{p_i - p_A}{nRel(A) + nRel(B)} - (q_i - q_A)\lambda_{A,B}\right]_+ \\
&= \left[\frac{p_i - p_A}{nRel(A) + nRel(B)} - \frac{(q_i - q_A)}{nRel(A) + nRel(B)}\left(\frac{p_A - p_B}{q_A + q_B}\right)\right]_+ \\
&= \frac{1}{nRel(A) + nRel(B)}\left[\frac{p_B(q_i - q_A) + q_B(p_i - p_A)}{q_A + q_B}\right]_+ \\
&= 0
\end{aligned}$$

The second last term evaluates to ≤ 0 and so the last equality holds because λ'_i is clipped to 0.

Thus, for any element not been selected by the EOR Algorithm i.e. $i > k_A$, the corresponding dual variable $\lambda'_i = 0$. Analogously, for any element $j > k_B$ in group B it can be shown that $\lambda'_j = 0$. We have shown that for any element not selected by the EOR Algorithm the corresponding dual variable $\lambda' = 0$.

Case II: Elements selected by the EOR Algorithm.

We have i) $p_i \geq p_A$ and $q_i \geq q_A$ as EOR selects in decreasing order of probabilities, and ii) $\lambda_{A,B} \geq 0$ or $\lambda_{B,A} \geq 0$ as only one of them can be non zero.

In Eq. (21), if $\lambda_{B,A} \geq 0$, then $\lambda_{A,B} = 0$ and with $p_i \geq p_A$, $q_i \geq q_A$ the resultant quantity in (21) would be ≥ 0, so that $\lambda'_i \geq 0$.

$$\begin{aligned}
\lambda'_i &= \left[\frac{p_i - p_A}{nRel(A) + nRel(B)} + (q_i - q_A)\lambda_{B,A}\right]_+ \\
&\geq 0
\end{aligned}$$

In Eq. (21), if $\lambda_{A,B} \geq 0$, then $\lambda_{B,A} = 0$. We can then substitute $\lambda_{A,B} = \frac{1}{nRel(A)+nRel(B)}\left(\frac{p_A - p_B}{q_A + q_B}\right)$ in (21),

$$\begin{aligned}
\lambda'_i &= \left[\frac{p_i - p_A}{nRel(A) + nRel(B)} - (q_i - q_A)\lambda_{A,B}\right]_+ \\
&= \left[\frac{p_i - p_A}{nRel(A) + nRel(B)} - \frac{(q_i - q_A)}{nRel(A) + nRel(B)}\left(\frac{p_A - p_B}{q_A + q_B}\right)\right]_+ \\
&= \frac{1}{nRel(A) + nRel(B)}\left[\frac{p_B(q_i - q_A) + q_B(p_i - p_A)}{q_A + q_B}\right]_+ \\
&\geq 0
\end{aligned}$$

The second last term evaluates to ≥ 0, so the last equality holds. Thus, for any element selected by the EOR Algorithm in group A i.e. $i \leq k_A$, the corresponding dual variable $\lambda' \geq 0$. Analogously, for any element $j \leq k_B$ in group B, $\lambda'_i \geq 0$. We have shown that for any element selected by the EOR Algorithm the corresponding dual variable $\lambda' \geq 0$.

We now show that $\lambda_k \geq 0$. From Eq. (16),

$$\lambda_k = \left[\frac{p_A}{nRel(A) + nRel(B)} - q_A(\lambda_{A,B} - \lambda_{B,A})\right] \tag{22}$$

If $\lambda_{A,B} \geq 0$, then $\lambda_{B,A} = 0$. Substituting $\lambda_{A,B} = \frac{1}{nRel(A)+nRel(B)}\left(\frac{p_A - p_B}{q_A + q_B}\right)$ in Eq. (22),

$$\begin{aligned}
\lambda_k &= \left[\frac{p_A}{nRel(A) + nRel(B)} - q_A\lambda_{A,B}\right] \\
&= \left[\frac{p_A}{nRel(A) + nRel(B)} - \frac{q_A}{nRel(A) + nRel(B)}\left(\frac{p_A - p_B}{q_A + q_B}\right)\right] \\
&= \frac{1}{nRel(A) + nRel(B)}\left(\frac{p_A q_B + q_A p_B}{q_A + q_B}\right) \\
&\geq 0
\end{aligned}$$

The last inequality follows since each of the terms p_A, q_A, p_B, q_B are ≥ 0. If $\lambda_{B,A} \geq 0$, then $\lambda_{A,B} = 0$. By substituting $\lambda_{B,A} = \frac{1}{nRel(A)+nRel(B)}\left(-\frac{p_A - p_B}{q_A + q_B}\right)$ in Eq. (22), we similarly get $\lambda_k \geq 0$.

The two duals $\lambda_{A,B}, \lambda_{B,A}$ are ≥ 0 by their construction in Eqs. (14), (15). Thus, we have shown that all the constructed dual variables $\lambda \geq 0$. □

Lemma 5.3. *The dual variables $\lambda = [\lambda'_1 \cdots \lambda'_n, \lambda_k, \lambda_{A,B}, \lambda_{B,A}]$ are always feasible.*

Proof. In Lemma 5.2, we proved that the constructed $\lambda \geq 0$. We now show that they satisfy the duality constraint.

For some element i, the duality constraint implies that

$$q_i(\lambda_{A,B} - \lambda_{B,A}) + \lambda_k + \lambda'_i \geq \frac{p_i}{nRel(A) + nRel(B)} \tag{23}$$

Without loss of generality, we consider element at index i that belongs to group A. Similar to Lemma 5.2, we consider two cases.

Case I: Elements not selected by the EOR Algorithm.

Using the fact that $\lambda'_i = 0$ for $i > k_A$ from Lemma 5.2, and substituting λ_k from Eq. (16), we get

$$\begin{aligned}
q_i(\lambda_{A,B} - \lambda_{B,A}) + \lambda_k + \lambda'_i &= q_i(\lambda_{A,B} - \lambda_{B,A}) + \lambda_k \\
&= q_i(\lambda_{A,B} - \lambda_{B,A}) + \frac{p_A}{nRel(A) + nRel(B)} - q_A(\lambda_{A,B} - \lambda_{B,A}) \\
&= \frac{p_A}{nRel(A) + nRel(B)} + (q_i - q_A)(\lambda_{A,B} - \lambda_{B,A})
\end{aligned}$$

We have i) $p_i \leq p_A$ and $q_i \leq q_A$ as EOR selects in decreasing order of probabilities, and ii) either $\lambda_{A,B} \geq 0$ or $\lambda_{B,A} \geq 0$ as only one of them can be nonzero. If $\lambda_{A,B} \geq 0$, then substituting $\lambda_{A,B}$,

$$\begin{aligned}
q_i(\lambda_{A,B} - \lambda_{B,A}) + \lambda_k + \lambda'_i &= \frac{p_A}{nRel(A) + nRel(B)} + (q_i - q_A)\lambda_{A,B} \\
&= \frac{p_A}{nRel(A) + nRel(B)} + \frac{(q_i - q_A)}{nRel(A) + nRel(B)}(\frac{p_A - p_B}{q_A + q_B}) \\
&= \frac{1}{nRel(A) + nRel(B)}\left[p_i + \frac{p_B(q_A - q_i) + q_B(p_A - p_i)}{q_A + q_B}\right] \\
&\geq \frac{p_i}{nRel(A) + nRel(B)}
\end{aligned}$$

Similarly, we can show that the dual constraint is satisfied if $\lambda_{B,A} \geq 0$. Thus, for any element not selected by the EOR Algorithm i.e. $i > k_A$, the corresponding dual constraint is satisfied. Analogously, for any element $j > k_B$ in group B it can be shown that the corresponding dual constraint is satisfied. We have shown that for any element not selected by EOR Algorithm the corresponding dual constraint is satisfied.

Case II: Elements selected by the EOR Algorithm.

Using the fact that $\lambda'_i \geq 0$ for $i \leq k_A$ from Lemma 5.2, and substituting λ_k from (16), λ'_i for $i \leq k_A$ in (23), we get

$$q_i(\lambda_{A,B} - \lambda_{B,A}) + \lambda_k + \lambda'_i = q_i(\lambda_{A,B} - \lambda_{B,A}) + \lambda_k + \frac{p_i}{nRel(A) + nRel(B)} - \lambda_k - q_i(\lambda_{A,B} - \lambda_{B,A}) = \frac{p_i}{nRel(A) + nRel(B)}$$

Thus, for any element selected by the EOR Algorithm i.e. $i \leq k_A$, $j \leq k_B$, the corresponding dual constraint is satisfied. □

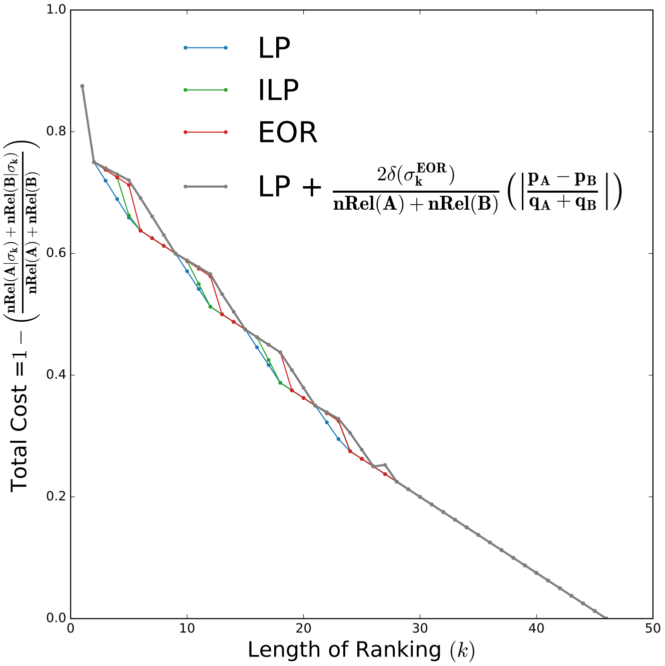


Figure 7: Cost Optimality Gap of a synthetic example with $p_{i\in A} = [1, 0.6, 0.5, 0.5, 0.4, 0.1, \cdots 0.1], nRel(A) = 4, S(A) = 15$, and $p_{i\in B} = [1, 0.1 \cdots 0.1], nRel(B) = 4, S(B) = 31$. The cost from EOR ranking is nearly optimal to the ILP or even the relaxed LP solution. Further the bound obtain in Theorem 5.1 (in grey) is tight for many k prefixes.

We demonstrate the cost optimality bound proved in Theorem 5.1 in Figure 7 that shows an example with a ranking produced by Linear Program (LP), Integer Linear Program (ILP), and the EOR algorithm along with the upper bound on the cost computed from the duality gap proved in Theorem 5.1. The example is constructed such that $\mathbb{P}(r_i|\mathcal{D})_{i\in A} = [1, 0.6, 0.5, 0.5, 0.4, 0.1, \cdots 0.1], nRel(A) = 4$, and $\mathbb{P}(r_i|\mathcal{D})_{i\in B} = [1, 0.1 \cdots 0.1], nRel(B) = 4$. Figure 7 shows that at most prefixes k, the EOR cost (in red) is optimal coinciding with the cost from ILP solution (in green) as well as with the LP solution (in blue). Further, when the EOR ranking does not coincide with the LP solution, the upper bound $\frac{2\delta(\sigma_k^{EOR})}{nRel(A)+nRel(B)}\left|\frac{p_A-p_B}{q_A+q_B}\right|$ is relatively small as is shown by the LP + duality gap (in grey).

We now present the proof for the global a priori bound on $\delta(\sigma_k^{EOR})$ for two groups A,B.

C.2 Proof for Theorem 5.2

Proof. Let $\sigma^{PRP,A}, \sigma^{PRP,B}$ be the PRP rankings for elements in group A and B respectively. We show by induction that for any given prefix k, EOR algorithm selects the element such that $\left|\delta(\sigma_k^{EOR})\right| \leq \delta_{max}$ and as a consequence of Theorem 5.1, we get a global cost guarantee of $\phi\delta_{max}$.

In the remaining proof, we drop the superscript of EOR for simplicity and σ_j refers to σ_j^{EOR}.

Consider the base case of $k = 1$. Algorithm 1 will select $\arg\min\left\{\frac{\sigma^{PRP,A}[1]}{nRel(A)}, \frac{\sigma^{PRP,B}[1]}{nRel(B)}\right\}$ resulting in the lower $\delta(\sigma_{k=1})$. If $\frac{\sigma^{PRP,A}[1]}{nRel(A)} \leq \frac{\sigma^{PRP,B}[1]}{nRel(B)}$, then $\delta(\sigma_1) = \frac{\sigma^{PRP,A}[1]}{nRel(A)} \leq \frac{1}{2}\left(\frac{\sigma^{PRP,A}[1]}{nRel(A)} + \frac{\sigma^{PRP,B}[1]}{nRel(B)}\right)$. Similarly, if $\frac{\sigma^{PRP,B}[1]}{nRel(B)} \leq \frac{\sigma^{PRP,A}[1]}{nRel(A)}$, then $\delta(\sigma_{k=1})$ denoted in short by $\delta(\sigma_1) = \frac{\sigma^{PRP,B}[1]}{nRel(B)} \leq \frac{1}{2}\left(\frac{\sigma^{PRP,A}[1]}{nRel(A)} + \frac{\sigma^{PRP,B}[1]}{nRel(B)}\right)$. Thus, at $k = 1$, by selecting the element with lower δ, EOR constraint is satisfied, i.e. $\delta(\sigma_1) \leq \delta_{max}$.

We assume that for a given $k-1$, $|\delta(\sigma_{k-1})| \leq \delta_{max}$. Further, without loss of generality, we assume that $\delta(\sigma_{k-1}) \geq 0$. We now show that at k, $|\delta(\sigma_k)| \leq \delta_{max}$ by considering the following cases. First, we show that if adding the element from one of the groups violates the δ_{max} constraint, then adding the element from the other group guarantees the satisfaction of δ_{max} constraint because EOR Algorithm selects the element that minimizes δ. Secondly, in the case where adding an element from either group does not violate the δ_{max} constraint, EOR algorithm will select the element that minimizes $|\delta(\sigma_k)|$ resulting in $|\delta(\sigma_k)| \leq \delta_{max}$. Finally, we show that when all the elements have run out from one of the groups at $k-1$, adding remaining elements from the other group will always satisfy the δ_{max} constraint.

We assume that adding the element from group A with relevance probability p_i at k, exceeds the δ_{max} constraint.

$$\delta(\sigma_{k-1}) + \frac{p_i}{nRel(A)} > \delta_{max} \tag{24}$$

Adding the element p_j from B at this prefix,

$$\delta(\sigma_k) = \delta(\sigma_{k-1}) - \frac{p_j}{nRel(B)} \leq \delta(\sigma_{k-1}) \leq \delta_{max} \tag{25}$$

The last inequality holds by the induction assumption at $k - 1$.

Further, since $\frac{p_j}{nRel(B)} \leq \frac{\sigma^{PRP,B}[1]}{nRel(B)}$, and $\delta_{max} = \frac{1}{2}\left(\frac{\sigma^{PRP,A}[1]}{nRel(A)} + \frac{\sigma^{PRP,B}[1]}{nRel(B)}\right)$, the above can be reduced to

$$\begin{aligned} \delta(\sigma_k) &= \delta(\sigma_{k-1}) - \frac{p_j}{nRel(B)} \geq \delta(\sigma_{k-1}) - \frac{\sigma^{PRP,B}[1]}{nRel(B)} \\ \delta(\sigma_k) &\geq \delta(\sigma_{k-1}) + \frac{\sigma^{PRP,A}[1]}{nRel(A)} - 2\delta_{max} \end{aligned} \tag{26}$$

Now using $\frac{p_i}{nRel(A)} \leq \frac{\sigma^{PRP,A}[1]}{nRel(A)}$, and Eq. (26) above,

$$\delta(\sigma_k) \geq \delta(\sigma_{k-1}) + \frac{\sigma^{PRP,A}[1]}{nRel(A)} - 2\delta_{max} \geq \delta(\sigma_{k-1}) + \frac{p_i}{nRel(A)} - 2\delta_{max} \tag{27}$$

Using Eqs. (27) and (24),

$$\delta(\sigma_k) \geq \delta(\sigma_{k-1}) + \frac{p_i}{nRel(A)} - 2\delta_{max} > -\delta_{max} \tag{28}$$

We have shown that, if $|\delta(\sigma_k)|$ exceeds δ_{max} by adding the element from group A (from (24)), then the element in group B will satisfy $|\delta(\sigma_k)| \leq \delta_{max}$ (from (25) and (28)). Since the EOR algorithm minimizes $|\delta(\sigma_k)|$, it will select the element from group B at prefix k rather than the element from group A. Thus, $|\delta(\sigma_k)| \leq \delta_{max}$ in this case.

Similarly, we can show that if $|\delta(\sigma_k)|$ exceeds δ_{max} by adding the element from group B, then adding the element from group A would result in $|\delta(\sigma_k)| \leq \delta_{max}$ and would be selected by the EOR algorithm at prefix k.

Finally, we consider the case where all the elements in a particular group have already been selected. Without loss of generality, let's assume that this is true with all the elements in group B added by prefix $k - 1$. We need to show that adding from the remaining elements in group A would still satisfy $|\delta| \leq \delta_{max}$ for the remaining prefixes.

From our assumption, $\frac{nRel(B|\sigma_{k-1})}{nRel(B)} = 1$ since all elements from group B were selected at prefix $k - 1$. From the inductive hypothesis $|\delta(\sigma_{k-1})| \leq \delta_{max}$,

$$|\delta(\sigma_{k-1})| = \left|\frac{nRel(A|\sigma_{k-1})}{nRel(A)} - \frac{nRel(B|\sigma_{k-1})}{nRel(B)}\right| \leq \delta_{max} \tag{29}$$

Since $\frac{nRel(A|\sigma_{k-1})}{nRel(A)} \leq 1$ as some elements remain in group A,

$$\delta(\sigma_{k-1}) = \frac{nRel(A|\sigma_{k-1})}{nRel(A)} - 1 \geq -\delta_{max} \tag{30}$$

After adding the element p_i from group A at prefix k and from (30),

$$\begin{aligned} \delta(\sigma_k) &= \delta(\sigma_{k-1}) + \frac{p_i}{nRel(A)} = \frac{nRel(A|\sigma_{k-1})}{nRel(A)} - 1 + \frac{p_i}{nRel(A)} \geq -\delta_{max} + \frac{p_i}{nRel(A)} \\ \delta(\sigma_k) &\geq -\delta_{max} \end{aligned} \tag{31}$$

Additionally, since $\frac{nRel(A|\sigma_n)}{nRel(A)} = 1$ implying $\frac{nRel(A|\sigma_k)}{nRel(A)} \leq 1$,

$$\delta(\sigma_k) = \frac{nRel(A|\sigma_k)}{nRel(A)} - 1 \leq 0 \tag{32}$$

From Eqs. (31) and (32), $-\delta_{max} \leq \delta(\sigma_k) \leq 0$ and thus EOR algorithm will add all the remaining elements from group A resulting in $|\delta(\sigma_k)| \leq \delta_{max}$. Analogously, it can be shown that if all the elements from group A had been added by prefix k, adding the next element from group B would satisfy $|\delta(\sigma_k)| \leq \delta_{max}$.

Thus, we have shown that Algorithm 1 provides rankings such that for any prefix k, $|\delta(\sigma_k)| \leq \delta_{max}$, where $\delta_{max} = \frac{1}{2}\left(\frac{\sigma^{PRP,A}[1]}{nRel(A)} + \frac{\sigma^{PRP,B}[1]}{nRel(B)}\right)$. As a consequence of this and Theorem 5.1, EOR rankings have total cost bounded by $\phi\delta_{max}$ for any prefix k of the ranking, where $\phi = \frac{2}{nRel(A)+nRel(B)}\left|\frac{p_A - p_B}{q_A + q_B}\right|$.

□

Next, we present the proof comparing costs from π^{EOR}, π^{unif} at prefix k, where $\delta(\sigma_k^{EOR}) = 0$.

C.3 Proof for Proposition 5.1

Proof. When $\delta(\sigma_k^{EOR}) = 0$, by the definition of EOR fairness, we have that $\frac{nRel(A|\sigma_k^{EOR})}{nRel(A)} = \frac{nRel(B|\sigma_k^{EOR})}{nRel(B)}$. As a result, the total cost $(1 - \frac{nRel(A|\sigma_k^{EOR})+nRel(B|\sigma_k^{EOR})}{nRel(A)+nRel(B)})$ as well as subgroup cost would be equal to

$$1 - \frac{nRel(A|\sigma_k^{EOR})}{nRel(A)} = 1 - \frac{nRel(B|\sigma_k^{EOR})}{nRel(B)} \tag{33}$$

We also know that $\frac{nRel(A|\sigma_k^{EOR})}{k_A} \geq \frac{nRel(A)}{S(A)}$ and $\frac{nRel(B|\sigma_k^{EOR})}{k_B} \geq \frac{nRel(B)}{S(B)}$, since the EOR algorithm selects top k_A, k_B elements from each of the groups (with $k_A + k_B = k, S(A) + S(B) = n$), having a higher mean relevance than that of the group itself.

$$\frac{nRel(A|\sigma_k^{EOR})S(A)}{nRel(A)} \geq k_A \tag{34}$$

$$\frac{nRel(B|\sigma_k^{EOR})S(B)}{nRel(B)} \geq k_B \tag{35}$$

Adding Eqs. (34), (35) and using (33), we get that

$$\begin{aligned} \frac{nRel(A|\sigma_k^{EOR})(S(A)+S(B))}{nRel(A)} &\geq k \\ \frac{nRel(A|\sigma_k^{EOR})}{nRel(A)} &\geq \frac{k}{n} \Leftrightarrow 1 - \frac{nRel(A|\sigma_k^{EOR})}{nRel(A)} \leq 1 - \frac{k}{n} \end{aligned}$$

This and Eq. (33) are sufficient to claim that the total cost and subgroup costs of uniform policy given by $1 - \frac{k}{n}$ will always be higher than the total cost and subgroup costs given by EOR ranking when $\delta(\sigma_k^{EOR}) = 0$. □

D Extension to Multiple Groups G

In the following, we prove the global cost and fairness guarantee for multiple groups G.

D.1 Proof for Theorem 6.1

Proof. The overall strategy for this proof is to consider each pair of groups among the $\frac{G(G-1)}{2}$ pairs and reduce each term of the duality gap to the two group case in Theorem 5.1. Fortunately, we can achieve such a reduction by careful construction of the dual variables.

The LP to find a solution X for this problem is formulated as follows

$$\max_{x \geq 0} \quad f(x) = \frac{P^T X}{\sum_{g=1}^{G} nRel(g)} \tag{Primal}$$

$$\text{s.t.} \quad X \leq 1 \tag{36}$$

$$X^T.1 \leq k \tag{select up to k elements}$$

$$G(G-1) \text{ constraints} \begin{cases} Q_{A,B}^T X & \leq & \delta(\sigma_k^{EOR}) \\ Q_{B,A}^T X & \leq & \delta(\sigma_k^{EOR}) \\ & \vdots & \end{cases}$$

The above LP is analogous to the two group case in Theorem 5.1, with the addition of $G(G-1)$ pairwise constraints ensuring EOR-fairness for all pairs of groups.

We can construct the dual problem as follows

$$\min_{\lambda \geq 0} \quad g(\lambda) = \delta(\sigma_k^{EOR}) \overbrace{\sum_{\{A,B\}} (\lambda_{A,B} + \lambda_{B,A})}^{G(G-1)/2} + k\lambda_k + \sum_{i=1}^{n} \lambda_i' \tag{Dual}$$

$$\text{s.t.} \quad \sum_{\{A,B\}} Q_{A,B}(\lambda_{A,B} - \lambda_{B,A}) + \lambda_k + \lambda' \geq \frac{P}{\sum_g nRel(g)} \tag{37}$$

We have pairs of dual variables that are constructed from the EOR solution as following

$$\lambda_{A,B} = \frac{1}{(G-1)}\frac{1}{\sum_g nRel(g)}\left[\frac{p_A - p_B}{q_A + q_B}\right]_+ \tag{38}$$

$$\lambda_{B,A} = \frac{1}{(G-1)}\frac{1}{\sum_g nRel(g)}\left[-\left(\frac{p_A - p_B}{q_A + q_B}\right)\right]_+ \tag{39}$$

$$\vdots$$

$$G(G-1) \quad \lambda' s$$

We construct λ_i' corresponding to constraint (36) and λ_k corresponding to constraint (select up to k elements) below.

$$\lambda_k = \left[\frac{p_A}{\sum_g nRel(g)} - q_A \overbrace{\sum_{g\neq A}(\lambda_{A,g} - \lambda_{g,A})}^{(G-1)\text{terms}}\right] = \left[\frac{p_B}{\sum_g nRel(g)} - q_B \overbrace{\sum_{g\neq B}(\lambda_{B,g} - \lambda_{g,B})}^{(G-1)\text{terms}}\right] = \cdots \text{for each of } G \text{ groups} \tag{40}$$

$$\lambda'_{i\in g'} = \left[\frac{p_i}{\sum_g nRel(g)} - \lambda_k - q_i \overbrace{\sum_{g\neq g'}(\lambda_{g',g} - \lambda_{g,g'})}^{(G-1)\text{terms}}\right]_+ \tag{41}$$

For instance, if $i \in A$ then,

$$\lambda'_{i\in A} = \left[\frac{p_i}{\sum_g nRel(g)} - \lambda_k - q_i \sum_{g\neq A}(\lambda_{A,g} - \lambda_{g,A})\right]_+$$

We show that the constructed dual variables are non-negative in Lemma 6.2 and always feasible in Lemma 6.3. Additionally, we have $\lambda' = 0$ for any element not selected in the EOR ranking from Lemma 6.2.

The duality gap can now be formulated as follows

$$g(\lambda^*) - f(X) = \delta(\sigma_k^{EOR}) \sum_{\{A,B\}} (\lambda_{A,B} + \lambda_{B,A}) + k\lambda_k + \sum_{i=1}^{n} \lambda_i' - \frac{P^T X}{\sum_g nRel(g)}$$

Substituting the values for λ' from (41) and breaking the k elements selected into k_A from group A, k_B from group B, and so on from every group, we have the above duality gap as

$$= \delta(\sigma_k^{EOR}) \sum_{\{A,B\}} (\lambda_{A,B} + \lambda_{B,A}) + k\lambda_k + \left(\overbrace{\sum_{i=1}^{k_A}\left(\frac{p_i}{\sum_g nRel(g)} - \lambda_k - q_i \sum_{g\neq A}(\lambda_{A,g} - \lambda_{g,A})\right) + \sum^{k_B}(.) + \cdots}^{G \text{ terms, one for each group}}\right) - \frac{P^T X}{\sum_g nRel(g)}$$

In the above G terms, we can collect $\sum^{k_A} \lambda_k + \sum^{k_B} \lambda_k + \cdots = k\lambda_k$ and $\left(\sum^{k_A} \frac{p_i}{\sum_g nRel(g)} + \sum^{k_B} \frac{p_j}{\sum_g nRel(g)} + \cdots\right) = \frac{P^T X}{\sum_g nRel(g)}$. This reduces the duality gap to

$$\begin{aligned} &= \delta(\sigma_k^{EOR})(\sum_{\{A,B\}} (\lambda_{A,B} + \lambda_{B,A})) \overbrace{- \sum_{i=1}^{k_A} q_i \sum_{g\neq A}(\lambda_{A,g} - \lambda_{g,A}) - \sum_{j=1}^{k_B} q_j \sum_{g\neq B}(\lambda_{B,g} - \lambda_{g,B}) - \cdots}^{G \text{ terms}} \\ &= \sum_{\{A,B\}} \left(\delta(\sigma_k^{EOR})(\lambda_{A,B} + \lambda_{B,A}) - (\sum_{i=1}^{k_A} q_i - \sum_{j=1}^{k_B} q_j)(\lambda_{A,B} - \lambda_{B,A})\right) \end{aligned}$$

For each pair of groups A, B, the term inside the summation reduces to the two group case in Theorem 5.1. We also have that $-\delta(\sigma_k^{EOR}) \leq \sum^{k_A} q_i - \sum^{k_B} q_j \leq \delta(\sigma_k^{EOR})$ from Lemma 6.1.

$$\begin{aligned}\text{Duality gap} &\leq \sum_{\{A,B\}} 2\lambda_{A,B}\delta(\sigma_k^{EOR}) \\ &\leq \frac{2\delta(\sigma_k^{EOR})}{(G-1)\sum_g nRel(g)} \sum_{\{A,B\}} \left|\frac{p_A - p_B}{q_A + q_B}\right|\end{aligned}$$

This proves that the EOR solution can only be ever as worse as $\phi\delta_{max}$ when compared with the optimal solution, where $\phi = \frac{2}{(G-1)\sum_{g=1}^{G} nRel(g)}\left(\sum_{\{A,B\}}\left|\frac{p_A-p_B}{q_A+q_B}\right|\right)$ and $\delta_{max} = \max_g\left\{\frac{\sigma^{PRP,g}[1]}{nRel(g)}\right\}$ from Lemma 6.4. □

Lemma 6.1. EOR ranking is $\delta(\sigma_k^{EOR})$ fairness optimal, implying that for all G choose 2 possible pairs of groups $A, B \in \{1, \cdots G\}$, we have $-\delta(\sigma_k^{EOR}) \leq \sum_{i=i}^{k_A} q_i - \sum_{j=1}^{k_B} q_j \leq \delta(\sigma_k^{EOR})$.

This lemma follows directly from the EOR ranking principle of choosing the candidate that minimizes $\delta(\sigma_k^{EOR})$ defined according to Eq. (8).

□

Lemma 6.2. The constructed dual variables $\lambda \geq 0$. In particular, for any $i > k_g$ in group g, where $g \in \{1, \cdots G\}$, it holds that $\lambda'_i = 0$ and for any $i \leq k_g$ it holds that $\lambda'_i \geq 0$.

Proof. In this Lemma, we show that $\lambda' = 0$ for the elements not selected and $\lambda' \geq 0$ for the selected elements by the EOR Algorithm. Without loss of generality, we consider the element at index i that belongs to group A.

$$\begin{aligned}\lambda'_{i\in A} &= \left[\frac{p_i}{\sum_g nRel(g)} - \lambda_k - q_i \sum_{g\neq A}(\lambda_{A,g} - \lambda_{g,A})\right]_+ \\ &= \left[\frac{p_i}{\sum_g nRel(g)} - \frac{p_A}{\sum_g nRel(g)} + q_A \sum_{g\neq A}(\lambda_{A,g} - \lambda_{g,A}) - q_i \sum_{g\neq A}(\lambda_{A,g} - \lambda_{g,A})\right]_+ \\ &= \left[\frac{p_i - p_A}{\sum_g nRel(g)} + (q_i - q_A)\sum_{g\neq A}(\lambda_{g,A} - \lambda_{A,g})\right]_+ \\ &= \left[\sum_{g\neq A}\left(\frac{p_i - p_A}{(G-1)\sum_g nRel(g)} + (q_i - q_A)(\lambda_{g,A} - \lambda_{A,g})\right)\right]_+ \end{aligned} \tag{42}$$

For every pair of $\lambda_{A,g}$ and $\lambda_{g,A}$, where $g \in \{1, \cdots G\}$ and $g \neq A$, only one of $\lambda_{A,g}, \lambda_{g,A}$ is ≥ 0. Each of the $G-1$ terms inside the summation in Eq. (42) reduces to the two group case as follows. For $i > k_A$ and each $\{A, g\}$, the term evaluates to ≤ 0 using Lemma 5.2 and thus λ'_i is clipped to 0. Similarly, for $i \leq k_A$ and each $\{A, g\}$ the term evaluates to ≥ 0 and thus $\lambda'_i \geq 0$.

We have shown that for any element not selected by EOR Algorithm the corresponding dual variable $\lambda' = 0$, and for any element selected by the EOR Algorithm the corresponding dual variable $\lambda' \geq 0$.

We now show that $\lambda_k \geq 0$. From Eq. (40),

$$\begin{aligned}\lambda_k &= \left[\frac{p_A}{\sum_g nRel(g)} - q_A \sum_{g\neq A}(\lambda_{A,g} - \lambda_{g,A})\right] \\ &= \sum_{g\neq A}\left(\frac{p_A}{(G-1)\sum_g nRel(g)} + q_A(\lambda_{g,A} - \lambda_{A,g})\right)\end{aligned} \tag{43}$$

Each of the $G-1$ terms inside the summation in Eq. (43) reduces to the two group case. For each $\{A, g\}$, the term evaluates to ≥ 0 using Lemma 5.2 and thus $\lambda_k \geq 0$.

The $G(G-1)$ duals $\lambda_{A,B}$ are ≥ 0 by their construction in (38). Thus, we have shown that all the constructed dual variables $\lambda \geq 0$. □

Lemma 6.3. The dual variables $\lambda = [\lambda'_1 \cdots \lambda'_n, \lambda_k, \lambda_{A,B}, \lambda_{B,A}, \cdots]$ are always feasible.

Proof. For some element $i \in A$, the duality constraint implies that

$$q_i \left(\overbrace{\sum_g (\lambda_{A,g} - \lambda_{g,A})}^{G-1 \text{ terms}} \right) + \lambda_k + \lambda'_i \geq \frac{p_i}{\sum_g nRel(g)} \tag{44}$$

Without loss of generality, we consider element $i \in A$.

Case I: Elements not selected by the EOR Algorithm.

Using the fact that $\lambda'_i = 0$ for $i > k_A$ from Lemma 6.2, and substituting λ_k from Eq. (40), we get

$$\begin{aligned} q_i \sum_{g \neq A} (\lambda_{A,g} - \lambda_{g,A}) + \lambda_k + \lambda'_i &= q_i \sum_{g \neq A} (\lambda_{A,g} - \lambda_{g,A}) + \frac{p_A}{\sum_g nRel(g)} - q_A \sum_{g \neq A} (\lambda_{A,g} - \lambda_{g,A}) \\ &= \frac{p_A}{\sum_g nRel(g)} + (q_i - q_A) \sum_{g \neq A} (\lambda_{A,g} - \lambda_{g,A}) \\ &= \sum_{g \neq A} \left(\frac{p_A}{(G-1) \sum_g nRel(g)} + (q_i - q_A)(\lambda_{A,g} - \lambda_{g,A}) \right) \qquad (45) \\ &\geq \sum_{g \neq A} \frac{p_i}{(G-1) \sum_g nRel(g)} \geq \frac{p_i}{\sum_g nRel(g)} \qquad (46) \end{aligned}$$

Each of the $G-1$ terms inside the summation in Eq. (45) reduces to the two group case. For each $\{A, g\}$, the term evaluates to $\frac{p_i}{(G-1) \sum_g nRel(g)}$ using Lemma 5.3 and thus the corresponding duality constraint is satisfied.

Case II: Elements selected by the EOR Algorithm.

Using the fact that $\lambda'_i \geq 0$ for $i \leq k_A$ from Lemma 6.2, and substituting λ_k from Eq. (40), λ'_i for $i \leq k_A$ in (41), we get

$$\begin{aligned} q_i \sum_{g \neq A} (\lambda_{A,g} - \lambda_{g,A}) + \lambda_k + \lambda'_i &= q_i \sum_{g \neq A} (\lambda_{A,g} - \lambda_{g,A}) + \lambda_k + \frac{p_A}{\sum_g nRel(g)} - \lambda_k - q_i \sum_{g \neq A} (\lambda_{A,g} - \lambda_{g,A}) \\ &= \frac{p_A}{\sum_g nRel(g)} \geq \frac{p_i}{\sum_g nRel(g)} \end{aligned}$$

Thus, for elements selected by the EOR Algorithm i.e. $i \leq k_A$, the corresponding dual constraint is satisfied. □

We now present the proof for the global a priori bound on $\delta(\sigma_k^{EOR})$ for G groups.

Lemma 6.4. The global a priori bound on $\delta(\sigma_k^{EOR})$ for G groups is given by $\delta_{max} = \max_g \left\{ \frac{\sigma^{PRP,g}[1]}{nRel(g)} \right\}$

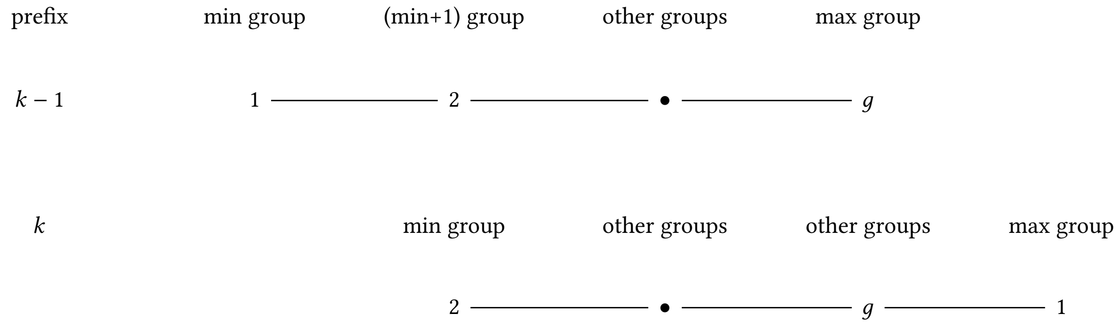


Figure 8: Illustration for the case of Multiple groups

Proof. We will show that for G groups, the value of δ_{max} such that a feasible ranking will be provided and that always satisfies $\delta(\sigma_k^{EOR}) \leq \delta_{max}$ for every given k is given by

$$\delta_{max} = \max \left(\frac{\sigma^{PRP,1}[1]}{nRel(1)}, \frac{\sigma^{PRP,2}[1]}{nRel(2)}, \cdots, \frac{\sigma^{PRP,g}[1]}{nRel(g)} \cdots \frac{\sigma^{PRP,G}[1]}{nRel(G)} \right) \tag{47}$$

In the remaining, we drop the superscript of EOR for simplicity and σ_j refers to σ_j^{EOR}.

We argue by an inductive argument similar to the proof of Theorem 5.2. Consider the base case of $k = 1$, when the first element is to be selected. The EOR algorithm will select according to Eq. (8) resulting in the lower $\delta(\sigma_{k=1})$. Thus, $\delta(\sigma_{k=1})$ is clearly $\le \delta_{max}$.

We assume that for a given $k-1$, $\delta(\sigma_{k-1}) \le \delta_{max}$ and show that at k, $\delta(\sigma_k) \le \delta_{max}$.

Consider the general case as depicted in Figure 8, where a group 1 has the lowest accumulated proportion and group g has the highest at prefix $k-1$. Since $\delta(\sigma_{k-1}) \le \delta_{max}$ from inductive assumption, we have

$$\frac{nRel(g|\sigma_{k-1})}{nRel(g)} - \frac{nRel(1|\sigma_{k-1})}{nRel(1)} \le \delta_{max}$$

At the next prefix k, if the group that is selected has $\frac{nRel(.|\sigma_k)}{nRel(.)} \le \frac{nRel(g|\sigma_{k-1})}{nRel(g)}$, then $\delta(\sigma_k) \le \delta_{max}$. Note that $\delta(\sigma_k)$ is always non-negative by definition from Eq. (8).

We now consider the case when a group g' is selected at the next prefix k such that $\frac{nRel(g'|\sigma_k)}{nRel(g')} > \frac{nRel(g|\sigma_{k-1})}{nRel(g)}$. Let us first consider that g' is group 1. We have $\frac{nRel(1|\sigma_k)}{nRel(1)} > \frac{nRel(g|\sigma_k)}{nRel(g)}$. Selecting group 1 at k means that the rest of the groups have the same accumulated relevance proportion $\frac{nRel(.|\sigma_k)}{nRel(.)}$ at prefix k as $k-1$. We analyze the difference of $\frac{nRel(.|\sigma_k)}{nRel(.)}$ between the group that was most behind– group 1 and the group that was second most behind – group 2 and whether that remains within δ_{max}. If the added element from group 1 is denoted by p_i, the EOR constraint value at k is

$$\begin{aligned}\delta(\sigma_k) &= \frac{nRel(1|\sigma_{k-1})}{nRel(1)} + \frac{p_i}{nRel(1)} - \frac{nRel(2|\sigma_k)}{nRel(2)} \\ &= \frac{p_i}{nRel(1)} - \left(\frac{nRel(2|\sigma_k)}{nRel(2)} - \frac{nRel(1|\sigma_{k-1})}{nRel(1)}\right) = \frac{p_i}{nRel(1)} - \left(\frac{nRel(2|\sigma_{k-1})}{nRel(2)} - \frac{nRel(1|\sigma_{k-1})}{nRel(1)}\right)\end{aligned} \tag{48}$$

Eq. (48) holds since group 1 is now the group with maximum relevance proportion after adding p_i - the top most current element from group 1. Group 2 becomes the group with minimum relevance proportion.

Since $\frac{p_i}{nRel(1)} \le \frac{\sigma^{PRP,1}[1]}{nRel(1)} \le \delta_{max}$ and because group 1 was behind group 2 at prefix $k-1$, we have $\frac{nRel(2|\sigma_{k-1})}{nRel(2)} \ge \frac{nRel(1|\sigma_{k-1})}{nRel(1)}$ since . As a result,

$$\delta(\sigma_k) \le \quad \frac{p_i}{nRel(1)} \le \frac{\sigma^{PRP,1}[1]}{nRel(1)} \le \delta_{max}$$

We have shown above that if the group with lowest relevance proportion at prefix $k-1$ (group 1 in this case) is selected and its relevance proportion now exceeds the group with the highest relevance proportion at prefix $k-1$ (group g in the case above), then $\delta(\sigma_k) \le \delta_{max}$. Thus, we can say that at least one group exists that satisfies δ_{max} EOR constraint at prefix k. This completes the proof that the EOR algorithm always provides a feasible ranking that satisfies $\delta_{max} = \max_{g \in \{1 \cdots G\}} \left\{\frac{\sigma^{PRP,g}[1]}{nRel(g)}\right\}$ for G groups. □

E Experiment Details

E.1 Baselines

We compare rankings from Algorithm 1 with the following baselines

Probability Ranking Principle (π^{PRP}). Candidates are selected in decreasing order of relevance independent of their group membership.

Uniform Policy (π^{unif}). Candidates are selected randomly independent of their group membership or relevance.

Thompson Sampling Ranking Policy (π^{TS}) [50]. For π^{TS}, binary relevances are drawn according to $r_i \sim \mathbb{P}(r_i|\mathcal{D})$, and candidates are sorted in decreasing order of relevance r_i with their ranking randomized for the same value of relevance r_i.

$$\pi^{TS} \sim \arg\operatorname{sort}_i[r_i] \quad s.t \quad r_i \sim \mathbb{P}(r_i|\mathcal{D})$$

π^{TS} ranks each candidate i in position k with probability that i has k^{th} highest relevance.

For both π^{TS} and π^{unif}, we compute expectation over 100 rankings $\sigma^{\text{unif}} \sim \pi^{\text{unif}}$ or $\sigma^{TS} \sim \pi^{TS}$ respectively and compute $\delta(\sigma_k)$ used in Table 1 as

$$\delta(\sigma_k) = \mathbb{E}_{\sigma\sim\pi}\left[\max_g \left\{\frac{nRel(g|\sigma_k)}{nRel(g)}\right\} - \min_g \left\{\frac{nRel(g|\sigma_k)}{nRel(g)}\right\}\right]$$

In order to plot a single ranking σ^{unif}, σ^{TS} for all experiments, we select the ranking with median $\sum_{k=1}^{n} |\delta(\sigma_k)|$

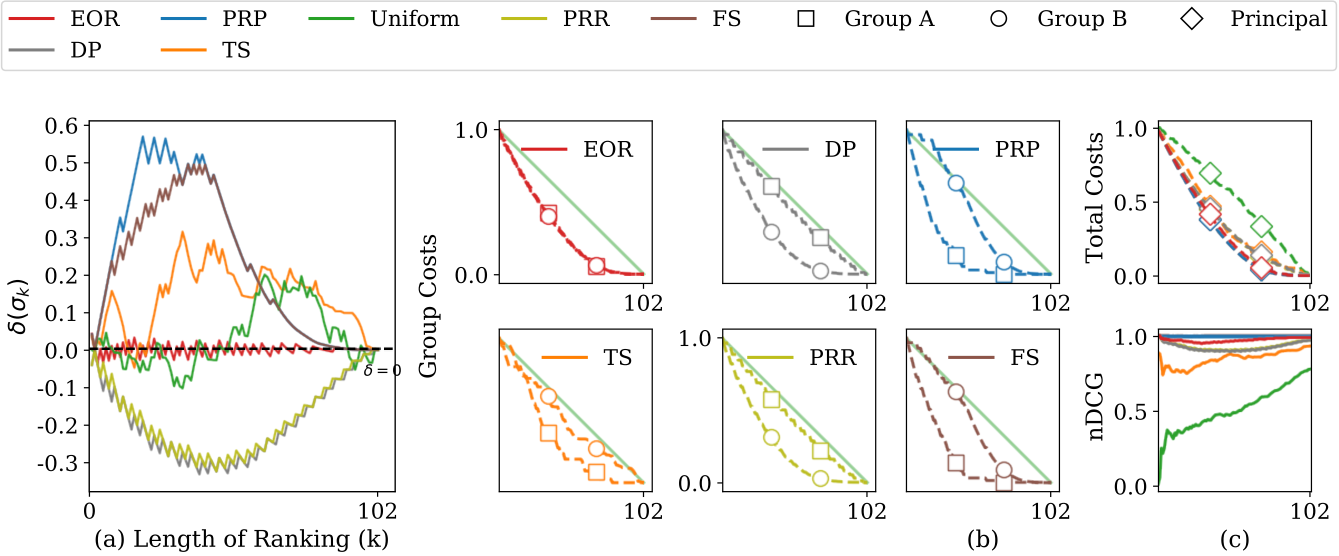


Figure 9: EOR criterion $\delta(\sigma_k)$, costs of the ranking policies, and DCG Utility for Synthetic dataset with proportional Rooney-Rule like constraint, π^{PRR}. For group A we draw $S(A) = 30$ relevance probabilities from $Powerlaw(\eta = 5)$, and then draw for group B from $Powerlaw(\eta = 0.5)$ until $nRel(A) \approx nRel(B)$.

Demographic Parity (π^{DP}). Candidates in each group are sorted in decreasing order of $\mathbb{P}(r_i|\mathcal{D})$ and selected such that the following constraint is minimized. This constraint is similar to the statistical parity variations introduced in [58].

$$\forall k \quad \frac{S(A|\sigma_k)}{S(A)} - \frac{S(B|\sigma_k)}{S(B)} \tag{49}$$

where $S(.)$ represents the size of the group. For a fair comparison with σ^{EOR}, we use Algorithm 1 and instead of minimizing Eq. (6), we minimize the above demographic parity constraint (49). We now discuss other variations of proportional representation constraints that have been introduced in prior literature [8–10]. Generally, these constraints require that the disadvantaged group selected is at least a specific proportion α of top k.

$$S(B|\sigma_k) \geq \alpha k \tag{50}$$

where $\alpha = \frac{S(B)}{S(A)+S(B)}$ and Eq. (50) is used as the fairness constraint while maximizing the utility to the principal. This type of representational constraint by definition requires the designation of a disadvantaged group. By designating B as the disadvantaged group, the constraint for proportional Rooney-Rule policy [47], which we denote by π^{PRR} is as follows

$$\forall k \quad \frac{S(B|\sigma_k)}{k} \geq \frac{S(B)}{S(A)+S(B)}$$

We empirically compare π^{PRR} baseline with other ranking policies in Figure 9 and as expected, find that it is similar to the baseline of π^{DP}, where π^{PRR} and π^{DP} almost overlap. Thus for a fair and analogous comparison with π^{EOR}, we use (49) as the π^{DP} baseline for all empirical evaluations. For more than two groups, we extend the DP baseline with the selection rule based on group size as follows. In particular,

$$\begin{aligned} \delta(\sigma^{DP}) &= \max_g \left\{\frac{S(g|\sigma_k)}{S(g)}\right\} - \min_g \left\{\frac{S(g|\sigma_k)}{S(g)}\right\} \\ l_g &= \sigma^{PRP,g}[1] \quad \forall g \in \{1 \cdot G\} \\ g^* &= \underset{g\in[1..G]}{\arg\min}\, \delta(\sigma^{DP} \cup \{l_g\}); \quad l_{g^*} = \sigma^{PRP,g^*}[1] \end{aligned} \tag{51}$$

*FA*IR Ranking Principle (π^{FS}).* This criterion is anchored on the principle that a top-k ranking is fair when the proportion of disadvantaged candidates selected doesn't fall far below a required minimum proportion p. This is formalized with a Binomial distribution, and a confidence level $(1 - \alpha)$. A function of the binomial cdf is computed apriori and is used as an input in the FA*IR Algorithm. Since Binomial(p=0.5,n) corresponds to a ranking where at each position, a candidate from either group is selected randomly, FA*IR is a "softened" version of demographic parity (DP). As a result, FA*IR is fundamentally different from Axiom 1 and Definition 4.1 derived from the uniform lottery fairness because, unlike DP, the uniform lottery is anchored on selecting an equal fraction of relevance from each group. Unlike π^{EOR}, π^{FS} is oblivious to the relevance distribution and thus cannot take disparate uncertainty into account. FA*IR also requires the normative designation of a disadvantaged group.

Consider the following example for top k=4 selection, with the probability of relevance for group A = [0.7, 0.7, 0.7, 0.7, 0.1, 0.1], group size = 6, relevant candidates = 3.0. Similarly, the probability of relevance for group B = [0.5, 0.5, 0.5, 0.5, 0.5, 0.5], group size = 6, relevant candidates = 3.0. The EOR Ranking for top-4 is [0.5, 0.7, 0.5, 0.7] with 2 candidates from group A, and 2 from group B, resulting in $\delta(\sigma_4^{EOR}) = 0.13$. The π^{FS} Algorithm with Binomial(p=0.5, n=12), k=4 and $\alpha = 0.1$ requires that at least 1 candidate be selected from the disadvantaged group while maximizing the utility to the principal. FA*IR ranking with group B as the disadvantaged group is $\sigma_4^{FS} = [0.7, 0.7, 0.7, 0.5]$. It selects 3 candidates from group A, and 1 from group B, resulting in $\delta(\sigma_4) = 0.53$. If instead group A is designated as the disadvantaged group, $\sigma^{FS} = [0.7, 0.7, 0.7, 0.7]$ with all candidates selected from group A, and none from group B, resulting in $\delta(\sigma_4^{FS}) = 0.93$. Note that for both FA*IR rankings, far fewer relevant candidates are chosen from group B, even though both groups have an equal number of relevant candidates in expectation.

In all the empirical evaluations in this paper, we assign group B as the minority group for π^{FS} and use the fairsearch core library [6] with default parameters of $\alpha = 0.1$.

Next, we discuss two exposure-based formulations π^{EXP} and π^{RA}.

Exposure-based Disparate Treatment (π^{EXP}). This policy enforces that the allocation of exposure to each group is proportional to their average utility. Specifically for two groups A and B,

$$\frac{\text{Exposure}(A|\Sigma)}{U(A)} = \frac{\text{Exposure}(B|\Sigma)}{U(B)}$$

where Σ is the doubly stochastic ranking matrix obtained from solving the Linear Program in [49]. For multiple groups, the above constraint is added for each pair of groups. $\text{Exposure}(g|\Sigma) = \frac{\Sigma_{i,j} v_j}{S(g)}$, $v_j = \frac{1}{\log(j+1)}$ for the j^{th} position, and $U(g) = \frac{\sum_{i \in g} p_i}{S(g)} = \frac{nRel(g)}{S(g)}$. In particular for two groups A, B, we solve the following LP [49]

$$\begin{aligned}
&\text{Maximize} && P^T \Sigma v && \text{utility to the principal} \\
&\text{subject to} && \mathbb{1}^T \Sigma = \mathbb{1}^T && \text{(sum of probabilities for each position)} \\
& && \Sigma \mathbb{1} = \mathbb{1} && \text{(sum of probabilities for each candidate)} \\
& && 0 \leq \Sigma_{i,j} \leq 1 && \text{(valid probability)} \\
& && \left(\frac{\mathbb{1}_{i \in A}}{nRel(A)} - \frac{\mathbb{1}_{j \in B}}{nRel(B)} \right) \Sigma v = 0 && \text{(exposure constraint)}
\end{aligned} \tag{52}$$

The group cost is computed as $\frac{\mathbb{1}_g P \Sigma}{nRel(g)}$, total cost as $\frac{P\Sigma}{\sum_g nRel(g)}$ and EOR criterion as $\max_g \left\{ \frac{\mathbb{1}_g P \Sigma}{nRel(g)} \right\} - \min_g \left\{ \frac{\mathbb{1}_g P \Sigma}{nRel(g)} \right\}$

Rank Aggregation w. proportional allocation of Exposure. For π^{RA}, we modify the baseline for fair rank aggregation in [7] as follows. In fair rank aggregation, all n candidates are ranked by m voters to achieve a ranking with maximum consensus accuracy, where consensus may be according to different aggregation methods while achieving fairness of exposure w.r.t groups. [7] proposes an algorithm that finds the consensus maximizing ranking and then swaps the candidates such that the equality of exposure is satisfied in that ranking. To adapt this baseline, we use the ranking from utility maximizing π^{PRP} as the consensus ranking and use the algorithm from [7] to swap elements in PRP ranking until the exposure constraint below is satisfied,

$$\frac{\min_g Exposure(g)}{\max_g Exposure(g)} \geq \text{threshold}$$

A threshold of 0.95 is used in experiments and on average over 100 runs, an exposure of $0.96 \pm 0.01, 0.96 \pm 0.00, 0.97 \pm 0.00$ is achieved for high, medium, and low levels of disparate uncertainty respectively in Table 1.

E.2 Synthetic Dataset

To simulate disparate uncertainty between groups, we draw $\mathbb{P}(r_i|\mathcal{D})$ directly from specific probability distributions as follows. For Group A, we obtain $p_i \sim Beta(\frac{1}{20}, \frac{1}{20})$ and keep them fixed. We simulate 100 runs and in each run, p_i for group B are sampled as follows until $nRel(B) \approx nRel(A)$ (total expected relevance for groups can only differ by 1.0).

- High Disparate Uncertainty: $Beta(5, 5)$
- Medium Disparate Uncertainty: $Beta(\frac{1}{2}, \frac{1}{2})$
- Low Disparate Uncertainty: $Beta(\frac{1}{20}, \frac{1}{20})$. Note that even when both groups are drawn from the same distribution, any sampled instance still contains some amount of disparate uncertainty.

[6] https://github.com/fair-search/fairsearch-fair-python

Results for unfairness and effectiveness of rankings are reported with standard error in Table 1 (left). The posterior distributions in Table 1 (right) uses 50 samples for each candidate in group A, while for group B, the number of samples increases from 10 to 30 to 50 as the setting changes from high to medium to low disparate uncertainty respectively.

To estimate $\mathbb{P}(i \in \sigma_k^\pi)$ for stochastic policies– π^{unif} and π^{TS}, we draw $d = 10^3$ Monte Carlo samples and compute Monte Carlo estimate according to (53).

$$1 - \mathbb{P}(i \in \sigma_k^\pi) = \frac{1}{d} \sum_{d} \mathbb{I}_{i \notin \sigma_k} \tag{53}$$

We compute the costs using $\mathbb{P}(r_i|\mathcal{D}), \mathbb{P}(i \in \sigma_k^\pi)$ according to Eqs. (3), and (4).

In Figure 11, we plot a random sample from Table 1 according to the generation process described above. We also qualitatively analyze a commonly used measure of utility to the principal, namely, the expected Normalized Discounted Cumulative Gain (nDCG), which according to our model is,

$$nDCG(\sigma_k) = \frac{DCG(\sigma_k)}{iDCG} = \frac{\sum_{i \in \sigma_k} v_i r_i}{\sum_{i \in \sigma_k^{Ideal}} v_i r_i}; \qquad \sigma^{Ideal} = \arg\text{sort}_i r_i$$

where $v_i = log_2 \frac{1}{(1+i)}$ for the i^{th} position. When true relevance labels are known, for instance in US Census experiments in Figure 14, $r_i \in \{0, 1\}$ consists of the true relevance labels, otherwise in synthetic experiments in Figure 11, $r_i \in [0, 1]$ consists of the calibrated $\mathbb{P}(r_i = 1|\mathcal{D})$.

As shown in Figure 10, the nDCG for EOR ranking is only slightly lower than the nDCG optimal PRP ranking and competitive with all other ranking policies. In all of these experiments, we confirm our findings that $\pi^{EOR}, \pi^{\text{unif}}$ distribute the subgroup and total costs evenly while other ranking policies π^{PRP}, π^{DP}, and π^{TS} place a high cost burden on one of the groups. Further, for π^{EOR}, the total cost to the principal and nDCG utility is close to the optimal (but unfair) total cost and utility of π^{PRP}, indicated by overlapping lines in subplots (c) of Figure 11.

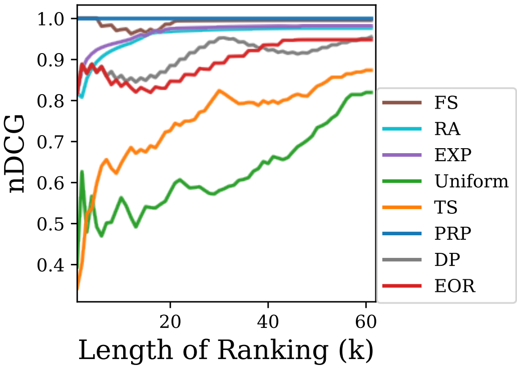


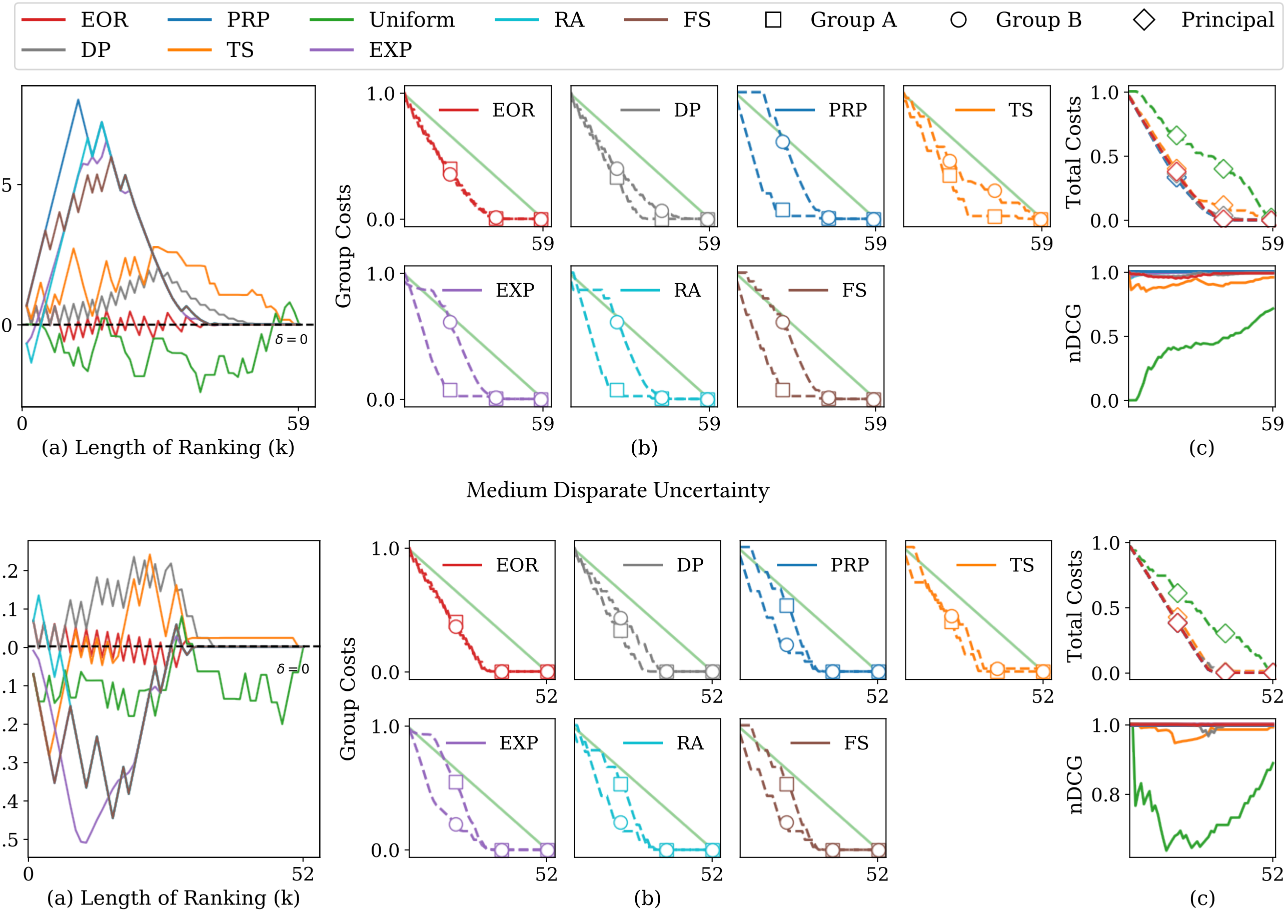
Figure 10: nDCG for High disparate uncertainty setting shown in Figure 4

E.3 US Census Survey Dataset

We use the ACSIncome task with default settings [14] for the state of New York and Alabama for 2018, with 1-year horizon. The dataset consists of 10 features, out of which 8 are categorical. Race is among the features that we include in the prediction task following [14]. There are 103,021 records for New York and 22,268 records for Alabama. For pre-processing, the categorical features are one-hot encoded, while the other two numerical features ('AGE' and 'WKHP') are standardized to have mean 0 and standard deviation 1. We divide this dataset into 60/20/20 for train/val/test split and fit a Gradient Boosting Classifier [7] with the parameters loss as 'exponential' and max_depth as 5 following hyperparameter configuration of [14]. This gives a DP violation $P(\hat{Y} = 1|White) - P(\hat{Y} = 1|Black)$ of 0.19 and an EO violation $P(\hat{Y} = 1|Y = 1, White) - P(\hat{Y} = 1|Y = 1, Black)$ of 0.18 for New York and a a DP violation of 0.22, EO violation of 0.29 for Alabama, which is roughly similar to Figure 2 and 6 of [14] before any fairness interventions are applied in the classification setting.

We subset the dataset to contain records with White or Black/African American racial membership (Alabama and New York) and subset records with White, Black, Asian, and Others racial membership (New York only) for two and four groups respectively. To calibrate relevance probabilities, we fit a Platt Scaling [37] calibrator on the validation data split group-wise and apply Platt Scaling to the test set probability estimates. Figure 12a, 12b and 12c show that calibrated $\mathbb{P}(r_i|\mathcal{D})$ on the test set, binned across 20 equal sized bins, lie close to the perfectly calibrated line.

[7] scikit-learn Gradient Boosting Classifier



(a) Length of Ranking (k)

(b)

(c)

Medium Disparate Uncertainty

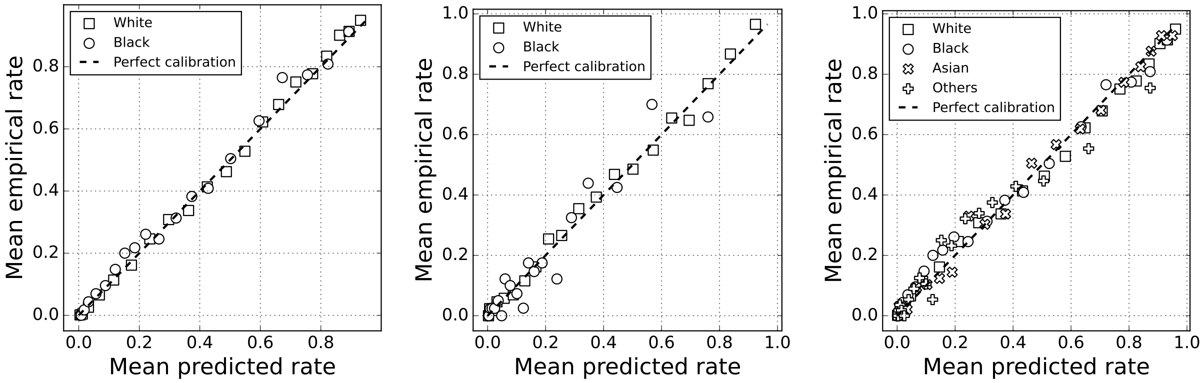
(a) Length of Ranking (k)

(b)

(c)

Low Disparate Uncertainty

Figure 11: Top: Medium disparate uncertainty Bottom Low disparate uncertainty for a randomly sampled instance.



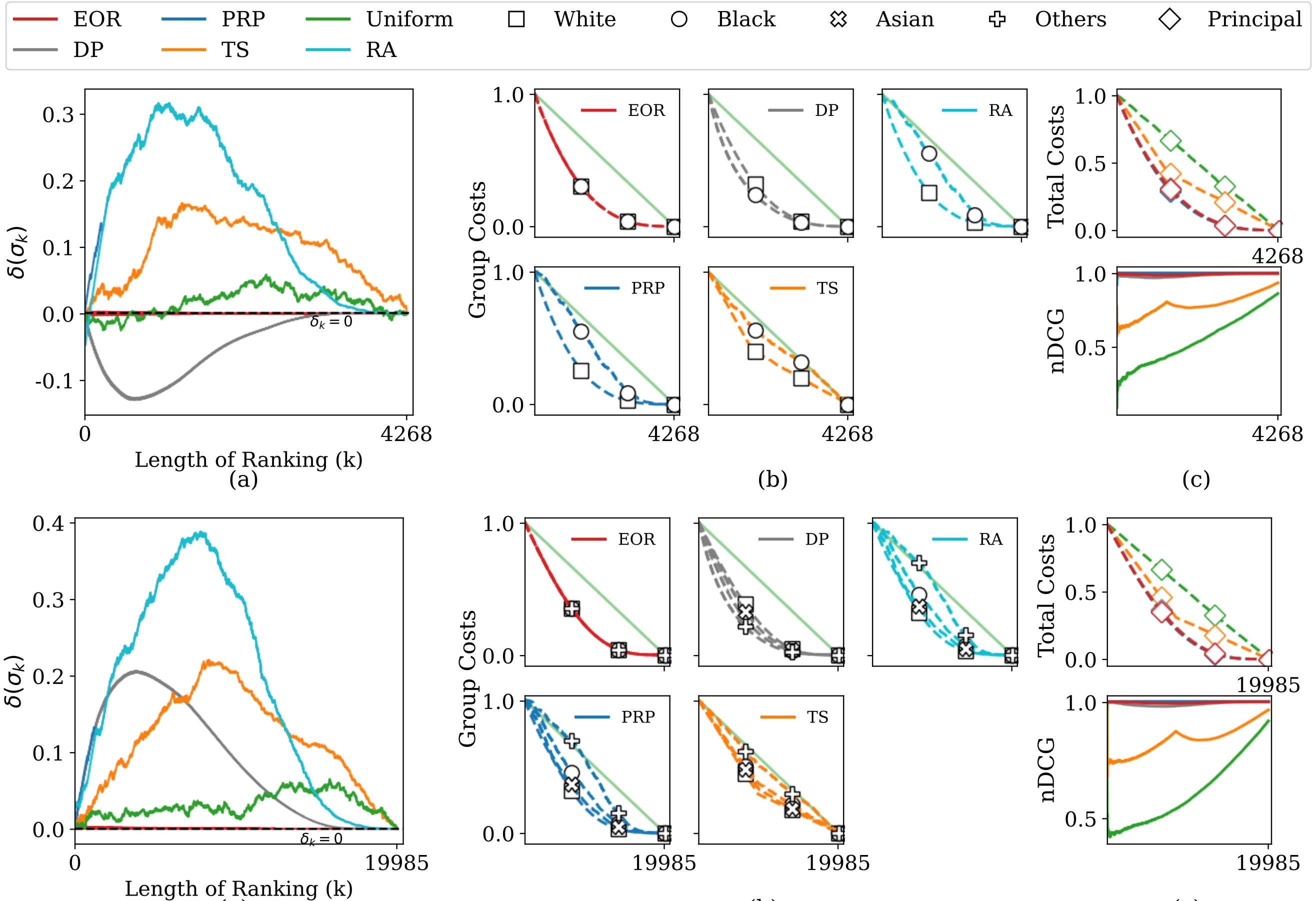
(a) New York

(b) Alabama

(c) New York

Figure 12: Calibration plot for $\mathbb{P}(r_i|\mathcal{D})$ for the state of New York and Alabama

In Figure 5, estimates $nRel(A|\sigma_k)$, $nRel(A)$, $nRel(B|\sigma_k)$, and $nRel(B)$ are computed with the true relevance labels from the test set for computing EOR criterion, costs, and nDCG. Figure 13, shows EOR criterion and costs with $nRel(A|\sigma_k)$, $nRel(A)$, $nRel(B|\sigma_k)$, $nRel(B)$ estimated



(a) (b) (c)

(a) (b) (c)

Figure 13: Top: EOR criterion $\delta(\sigma_k)$ and Costs computed using calibrated $\mathbb{P}(r_i|\mathcal{D})$ for two groups for the state of Alabama. **Bottom:** EOR criterion $\delta(\sigma_k)$ and Costs computed using calibrated $\mathbb{P}(r_i|\mathcal{D})$ for four groups for the state of New York.

from the calibrated $\mathbb{P}(r_i|\mathcal{D})$. Note that the evaluation on true relevance labels in Figure 5, though noisier is qualitatively similar to the evaluation using the calibrated $\mathbb{P}(r_i|\mathcal{D})$ in Figure 13. Additional experiment for two groups with true relevance labels for New York in Figure 14 (top) and with calibrated $\mathbb{P}(r_i|\mathcal{D})$ in Figure 14 (bottom) further confirm our findings, that π^{EOR} is the only ranking policy that consistently achieves $\delta(\sigma_k)$ close to zero at every prefix k with near optimal total cost to the principal.

Note the overlapping of π^{RA} and π^{PRP} in Figure 13 and 14. This is expected because π^{RA} swaps the candidates in PRP ranking to satisfy proportional exposure as described in Appendix E.1. Since the amortized exposure between groups is already satisfied with the PRP ranking for this dataset, π^{RA} and π^{PRP} compute similar rankings.

E.4 Amazon shopping queries dataset

Amazon's shopping queries [39] consists of a large scale query-product pair dataset with baseline models for tasks related to predicting the relevance of items given a search query. Each query-product pair has an associated human annotated label of an exact, substitute, complement, or irrelevant label.

For our analysis, we focus on their task 1 of query-product ranking [8] to sort the list of products in the decreasing order of relevance for every query. We use the publicly available baseline model for this task, consisting of Cross Encoders for the MS Marco dataset [40]. This pretrained model encodes the query and product titles and is fine-tuned on the US part of the small version of training dataset. We use the default hyperparameters for the Cross Encoder as maximum length=512, activation function=identity, and number of labels=1 (binary task). Similarly, for training following the default configuration, all exact labels are mapped to 1.0, while the rest (substitute, complement, and irrelevant) are mapped to 0.0. Default hyperparameter configuration includes MSE loss function, evaluation steps=5000, warm-up steps=5000, learning rate=7e-6, training epochs=1, and number of development queries=400. Inference from the trained model provides relevance scores and we apply a sigmoid function to transform these scores to probabilities of relevance $\mathbb{P}(r_i|\mathcal{D})$.

[8] https://github.com/amazon-science/esci-data

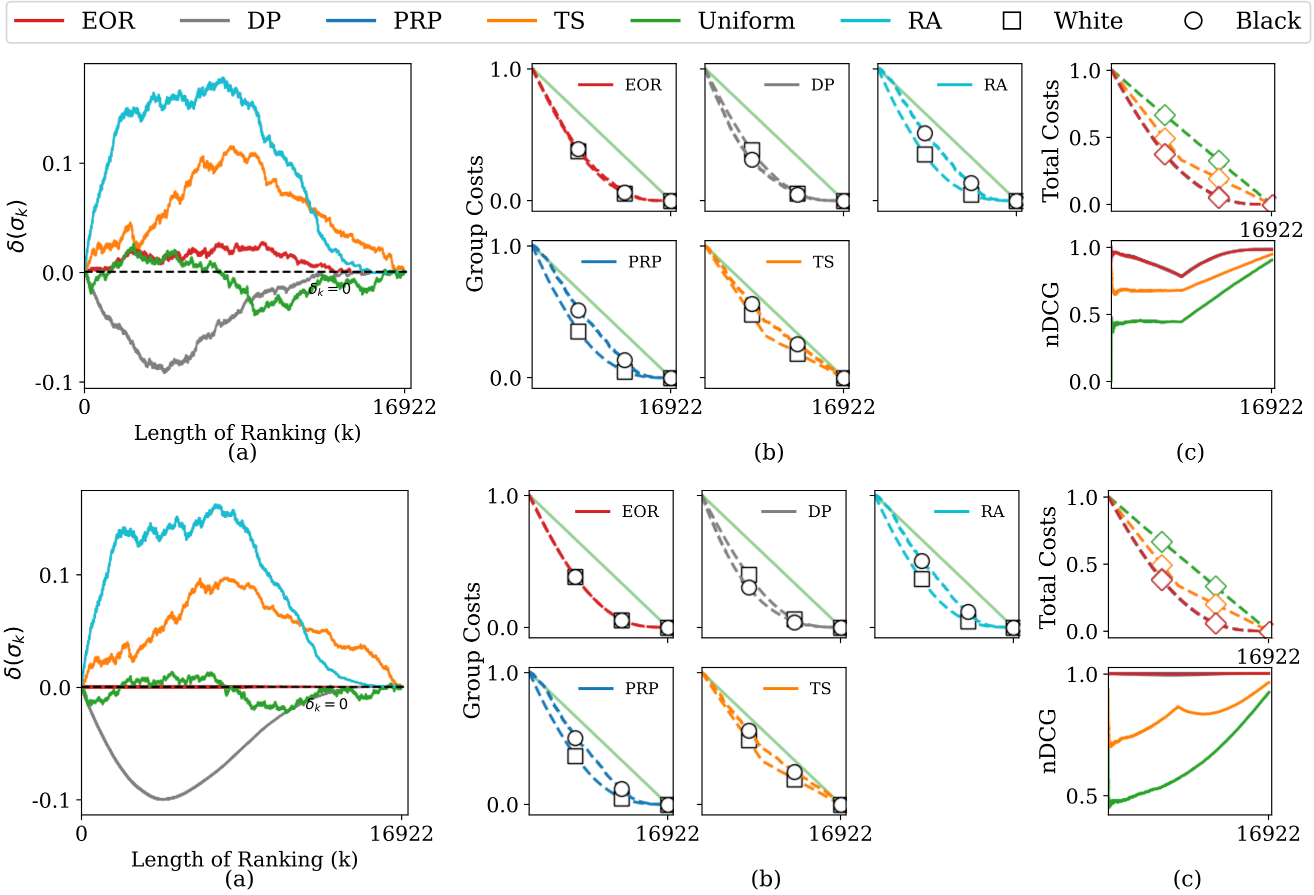


Figure 14: Top: EOR criterion $\delta(\sigma_k)$ and Costs computed using true relevance labels from the test subset. **Bottom:** EOR criterion $\delta(\sigma_k)$ and Costs computed using calibrated $\mathbb{P}(r_i|\mathcal{D})$ for the state of New York.

To evaluate the calibration of predicted $\mathbb{P}(r_i|\mathcal{D})$, we use the test split of the dataset [39] for the large version containing 22,458. We filtered these queries so that they contain at least three products owned by one of the 158 brands owned by Amazon (we discuss in the next paragraph the source of identifying these Amazon-owned brands) and at least three products owned by brands other than Amazon. These result in 395 queries, out of which half are used for calibration with a Platt-scaling calibrator while the remaining half is used to evaluate the calibration curve for the test dataset. $\mathbb{P}(r_i|\mathcal{D})$ of the query-product pairs for the remainder half of the test dataset after calibration is binned across 20 equal sized bins as shown in Figure 6a and lies close to the perfectly calibrated line.

We further augmented this with another dataset [9] collected from the Markup report [60], which investigated Amazon's placement of its own brand products as compared to other brands based on star ratings, reviews etc. The authors for the Markup report identified 158 brand products that are trademarked by Amazon. We use these 158 brands to form the Amazon owned group. Products belonging to any other brand form the non-Amazon group. Importantly, this dataset contains logged rankings from Amazon's website with 4566 queries for popularly searched query terms. We filtered these such that each query contains exactly 60 products and at least three of them are owned by Amazon, resulting in 1485 search queries.

Next, we obtain relevance probabilities $\mathbb{P}(r_i|\mathcal{D})$ from Amazon's pretrained baseline model described above and evaluate $\delta(\sigma_k)$ both for the logged ranking as well as our computed EOR ranking. Figure 6b shows that our EOR ranking is closer to $\delta(\sigma_k) = 0$ as compared to logged rankings on Amazon's platform. We note that this analysis is subject to confounding due to the use of features other than product titles that may be used in practice for logged rankings. However, the analysis does demonstrate how the EOR criterion can be used for auditing, if the auditor is given access to the production ranking model to avoid confounding.

[9] https://github.com/the-markup/investigation-amazon-brands