# Top-$K$ Contextual Bandits with Equity of Exposure

OLIVIER JEUNEN and BART GOETHALS, Adrem Data Lab, University of Antwerp, Belgium

The contextual bandit paradigm provides a general framework for decision-making under uncertainty. It is theoretically well-defined and well-studied, and many personalisation use-cases can be cast as a bandit learning problem. Because this allows for the direct optimisation of utility metrics that rely on online interventions (such as click-through-rate (CTR)), this framework has become an attractive choice to practitioners. Historically, the literature on this topic has focused on a one-sided, user-focused notion of utility, overall disregarding the perspective of content providers in online marketplaces (for example, musical artists on streaming services). If not properly taken into account – recommendation systems in such environments are known to lead to unfair distributions of attention and exposure, which can directly affect the income of the providers. Recent work has shed a light on this, and there is now a growing consensus that some notion of *"equity of exposure"* might be preferable to implement in many recommendation use-cases.

We study how the top-$K$ contextual bandit problem relates to issues of disparate exposure, and how this disparity can be minimised. The predominant approach in practice is to greedily rank the top-$K$ items according to their estimated utility, as this is optimal according to the well-known Probability Ranking Principle. Instead, we introduce a configurable tolerance parameter that defines an acceptable decrease in utility for a maximal increase in fairness of exposure. We propose a personalised exposure-aware arm selection algorithm that handles this relevance-fairness trade-off on a user-level, as recent work suggests that users' openness to randomisation may vary greatly over the global populace. Our model-agnostic algorithm deals with *arm selection* instead of utility modelling, and can therefore be implemented on top of any existing bandit system with minimal changes. We conclude with a case study on carousel personalisation in music recommendation: empirical observations highlight the effectiveness of our proposed method and show that exposure disparity can be significantly reduced with a negligible impact on user utility.

CCS Concepts: • **Information systems** → **Recommender systems**; • **Computing methodologies** → *Reinforcement learning*.

Additional Key Words and Phrases: Fairness; Probabilistic Models

**ACM Reference Format:**

Olivier Jeunen and Bart Goethals. 2021. Top-$K$ Contextual Bandits with Equity of Exposure. In *Fifteenth ACM Conference on Recommender Systems (RecSys '21), September 27–October 1, 2021, Amsterdam, Netherlands.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3460231.3474248

## 1 INTRODUCTION

This work investigates how the "*equity of exposure*" principle can be applied to top-$K$ contextual bandit problems for recommendation [3, 16]. Over the past few decades, recommendations have become an integral part of many platforms on the world wide web. They come in many shapes and sizes, but typically have a common focus: increasing engagement (in the form of clicks, streams, sales, subscriptions, …), by guiding users towards content that is relevant to them. In recent years, recommendation systems have come up in various scenarios where a one-sided, user-focused view of performance might be hurtful to other *stakeholders* in the system [1]. Indeed, many online platforms for music streaming, food delivery, ride-hailing or e-commerce can be seen as multi-stakeholder environments, where the platforms can benefit from implementing a notion of provider-side utility as well. Instead of merely focusing on increased engagement, the platform can then pursue various fairness-inspired goals such as "equity of attention" across content providers (the musical artists, restaurants, retailers and such on the platforms) [5].

This notion has recently been extended to the "equity of exposure" principle, which states that items of equal relevance should receive equal exposure in expectation [16]. Its importance is most palpable in multi-stakeholder platforms also known as *marketplaces*, as unfair *exposure* distributions can then directly lead to unfair *income* distributions [30].

The recommendation task can be interpreted in many ways, and various paradigms have been introduced over the years. A large body of work has focused on the "next-item prediction" task, where the system takes in sequences of *organic* user-item interactions and aims to predict which other items the user will interact with [40, 47]. The core problem with this framing is that it disregards any notion of *intervention* that the system takes by showing recommendations to users. As a result, offline evaluation results for next-item prediction are notoriously uncorrelated with typical online success metrics such as click-through-rate (CTR), and it is often unclear whether offline improvements translate to online gains [14, 18, 21, 42]. This in turn makes experimentation cycles in industry expensive, as the outcomes of online experiments are essentially unpredictable based on offline experiments alone.

Other works cast the recommendation task as a "contextual bandit" problem [3, 19, 23, 28]. This provides a general framework for decision-making under uncertainty, and allows recommendation researchers and practitioners to draw from the relevant literature on this well-studied learning paradigm [36, §8.3]. Bandit algorithms focus on maximising a utility function by performing interventions, learning from their own experience over time. Because this is much more closely tied to the recommendation problem as it occurs in practice (increasing engagement by showing the *right* recommendations), it has become an attractive choice for practitioners. Several specifics such as delayed feedback, top-$K$ recommendations and scalability requirements inhibit the straightforward adoption of bandit methods that were proposed in a general context, but many works have proven successful in bridging this gap [8, 11, 24, 28]. Most recently, Bendada et al. cast *carousel* personalisation in music streaming apps as a top-$K$ contextual bandit problem with delayed updates, reassuringly reporting alignment between offline and online evaluation results [3].

In this work, we study how the "equity of exposure" principle relates to top-$K$ contextual bandit problems in recommender systems. We propose a novel algorithm for *Exposure-Aware aRm Selection* (EARS), which tackles the relevance-fairness trade-off in a personalised manner. To the best of our knowledge, EARS is the first algorithm that deals with equity of exposure in top-$K$ contextual bandit recommenders. EARS is a general, model-agnostic algorithm. As a result, it can be applied to any existing bandit system with minimal changes. We validate EARS in a case-study on the real-world music streaming data released by Bendada et al. [3]. Empirical observations indicate that our algorithm is effective in minimising disparity of exposure with a minimal impact on expected rewards. All source code to reproduce our experimental results is available at github.com/olivierjeunen/EARS-recsys-2021.

## 2 RELATED WORK

As recommendation systems and machine learning applications in general become more widespread, there is a growing responsibility for researchers and practitioners to consider the impact that the systems they build have on the world around them. The importance of this notion of *fairness* in machine learning has been widely recognised in recent years, and several research areas now explicitly focus on it [2].

*Fairness in rankings.* In ranking systems such as search engines or recommenders, unfairness typically stems from uneven distributions of attention and exposure over the subjects that are being ranked. These distributions can either be measured on the individual level [5], or over groups defined by sensitive attributes like race or gender [44].

Several approaches have been proposed to directly optimise rankings for fairness – often by adapting Learning-To-Rank (LTR) problem formulations to explicitly include measures of individual- or group-fairness in the optimisation process [6, 35, 45, 50]. Focusing on more general information retrieval applications – Diaz et al. recently proposed the "*equity of exposure*" principle, exploiting the stochasticity of rankings to lead to fairness *in expectation* [16].

*Fairness in contextual bandits.* In contextual bandit scenarios, a notion of *meritocratic* fairness is often adopted [26]. This ensures that *better* arms are drawn with higher probabilities than *worse* arms, which aligns with the goals of equity of exposure. As we will discuss in Section 3.2, this still leads to deterministic and greedy arm selection when a total order is defined over the action space. Wang et al. consider fairness of exposure in contextual bandit scenarios as well, proposing to adopt a merit function which is used to ensure that exposure is proportional to merit [49]. Based on this, they propose a notion of *fairness regret* and prove regret bounds for Upper-Confidence-Bound (UCB) and Thompson Sampling-based algorithms to learn an optimal policy with respect to reward regret and fairness regret. In contrast, we specify a tolerance parameter that allows a maximal deviation from the reward obtained by the optimal policy, whilst maximising fairness. Their work deals with general, single-arm contextual bandits, whereas the focus of our work lies on ensuring fairness of exposure in top-$K$ recommendation problems. Additionally, we derive items' merit directly from their probability of leading to a positive reward, and deal with delayed model updates. Finally, Patil et al. lose the contextual component that is necessary to allow for personalisation [37], and Chen et al. do not consider meritocratic fairness, but rather ensure a minimal amount of exposure to *every* arm [12].

*Fairness in recommendation.* In recommender systems, fairness is often a multi-sided concept [7, 38, 48]. On one hand, there is fairness towards the *user* or *consumer*, in that similar users must be treated in similar ways (following the disparate treatment adage [2]). On the the other hand, there is the *item* or *provider* perspective, which is often linked to notions of equity of exposure, either individually measured or group-based. Deldjoo et al. propose a framework for evaluating both types of fairness in traditional recommendation algorithms [15] – departing from the contextual bandit formulation we adopt in this work. Beutel et al. propose a pairwise ranking method to improve group-fairness measures in recommendation [4]. Optimising for user-focused utility (in the form of relevance or clicks) is often unfavourable to item-fairness formulated as equity of exposure. Closely related to popularity bias, the myopic optimisation of one-sided metrics can lead to "superstar economics" and "winner-takes-all" dynamics, which is often undesirable. Mehrotra et al. present a conceptual framework for evaluating the trade-off between *relevance* and *fairness* that then naturally occurs, providing the main source of inspiration for this work [31]. Sonboli et al. also propose a user-level re-ranking method to increase provider fairness in recommendation, aimed at the classical recommendation task where additional user and item features are available [46]. Other related work adapts the optimisation problem at hand to include a notion of fairness or diversity into the objective function of the bandit algorithm, which originally focuses on maximising cumulative reward [27, 33]. In contrast, the method we propose in this work deals with arm selection instead of the modelling procedure itself. As a result, it can efficiently be implemented on top of existing systems with minimal changes, retaining the convenient convergence properties of state-of-the-art bandit methods.

# 3 METHODOLOGY AND CONTRIBUTIONS

## 3.1 Contextual Bandits for Top-$K$ Recommendation

The contextual bandit paradigm defines a modelling framework for decision-making under uncertainty. Every *round*, the system observes a context vector drawn from some unknown marginal distribution: $\boldsymbol{x} \sim \mathsf{P}(X)$. This context describes a

**Algorithm 1** Contextual Bandits for Top-$K$ Recommendation with Delayed Batch Updates

```
Input: Number of rounds $T$, number of opportunities per round $T_r$, number of slots $K$
Initialise model $\mathcal{M}$
for $t = 1, \dots, T$ do
    $\mathcal{D}_t = \emptyset$
    for $t_r = 1, \dots, T_r$ do
        Observe context $x_{t_r}$
        Select arms $a_{t_r,k} \forall k = 1, \dots, K$ according to the model $\mathcal{M}$
        Observe clicks $c_{t_r,k} \forall k = 1, \dots, K$
        $\mathcal{D}_t = \mathcal{D}_t \cup \{(x_{t_r}, a_{t_r}, c_{t_r})\}$
    Update model $\mathcal{M}$ with batch feedback $\mathcal{D}_t$
```

user visiting the system, and is often an embedded representation of their item consumption history together with contextual features describing the device they are using, the time of day, et cetera. In the top-$K$ recommendation setting, the system is prompted to select $K$ *arms* that correspond to items to show to the user, chosen from the set $\mathcal{A}$. Arms, items and actions are all synonymous in this context, referring to recommendations being shown to the user. A *policy* is a probability distribution over actions conditioned on contexts: $\pi(A|X) \equiv \mathsf{P}(A|X,\pi)$, describing which (lists of) recommendations are likely to be shown when the system is presented with context $x$.

We subsequently observe possible rewards $c \sim \mathsf{P}(C = 1|X = x; A = a)$ when the user interacts with the recommendations that we have shown. These rewards can consist of various different signals, but we will refer to them as *clicks* for the sake of notational brevity. After a number of opportunities have passed, the model responsible for selecting arms is updated with the context-action-reward triplets that were generated by the system since the last update. These are called "delayed batch updates", as there might be significant time between the opportunity and the update, and the new data is seen by the model in bulk. Algorithm 1 describes this iterative process of observing contexts, selecting arms, observing clicks and updating the model. Successful approaches to this learning setting balance the "*explore–exploit*" dilemma, as the objective at every round is both to (1) accumulate a high reward, and to (2) gather useful training data that improves the model at future rounds. Contextual bandit algorithms are typically evaluated in terms of *cumulative regret*, defined as the difference between the reward accumulated by their policy and the optimal oracle policy:

$$\pi^{\star}(A|x) = \arg\max_{\pi \in \Omega} \left( \mathop{\mathbb{E}}_{\mathcal{A}_K \sim \pi(A|x),\, x \sim \mathsf{P}(X)} \left[ \mathsf{P}(C = 1|X = x; A_{1:K} = \mathcal{A}_K) \right] \right). \tag{1}$$

Define $\mathcal{D}_{\pi,T}$ as the context-action-reward triplets collected under policy $\pi$ up to round $T$. The cumulative reward for a policy $\pi$ at round $T$ is the total number of clicks it has collected ($\sum\limits_{c \in \mathcal{D}_{\pi,T}} c$), and the cumulative regret can then be described as $\sum\limits_{c \in \mathcal{D}_{\pi^{\star},T}} c - \sum\limits_{c \in \mathcal{D}_{\pi,T}} c$. Various methods to elegantly handle the exploration–exploitation trade-off have been proposed over the years, some of which are provably optimal under mild assumptions. Upper-Confidence-Bound (UCB) [28] and Thompson Sampling (TS)-based methods [8, 17] make up the most common families of approaches.

It is common in the literature to adopt independent linear models for every arm [3, 28], but it should be noted that real-world applications can benefit from augmenting the bandit signal with organic user-item interactions to transfer bandit signal between arms [43]. Algorithm 2 describes the updating procedure for a general Online Bayesian Logistic Regression model that infers the probability of a click conditional on a context-action pair [8]. At every update, the Maximum A Posteriori (MAP) estimate for the model parameters $\boldsymbol{\theta}_a$ is computed with respect to the negative log-likelihood of the newly observed data and the model parameters at the previous iteration as prior. The

**Algorithm 2** Updating Independent Bayesian Logistic Regression Models

```
Input: Linear models with Gaussian priors for parameters $(\boldsymbol{\mu}_a, \boldsymbol{\sigma}_a) \forall a \in \mathcal{A}$, batch feedback $\mathcal{D}$ to update with
Output: Updated linear models with posterior parameters $(\boldsymbol{\mu}'_a, \boldsymbol{\sigma}'_a) \forall a \in \mathcal{A}$
for $a \in \mathcal{A}$ do
    $\mathcal{D}_a \coloneqq \{(x_i, a_i, c_i) \in \mathcal{D} \mid a_i = a\}$
    $\boldsymbol{\theta}_a = \arg\min_{\boldsymbol{\theta} \in \Theta} \left( \overbrace{\frac{1}{2} \left\| \frac{\boldsymbol{\theta} - \boldsymbol{\mu}_a}{\sqrt{\boldsymbol{\sigma}_a}} \right\|_2^2}^{\textbf{Prior}} + \overbrace{\sum_{(\boldsymbol{x},c) \in \mathcal{D}_a} \left( c \cdot \log(1 + \exp(-\boldsymbol{\theta}^\intercal \boldsymbol{x})) - (1 - c) \cdot \log(1 + \exp(-\boldsymbol{\theta}^\intercal \boldsymbol{x})) \right)}^{\textbf{Negative Log-Likelihood}} \right)$
    $\boldsymbol{\mu}'_a = \boldsymbol{\theta}_a; \boldsymbol{\sigma}'_a = \left( \boldsymbol{\sigma}_a^{-1} + \sum_{\boldsymbol{x} \in \mathcal{D}_a} \boldsymbol{x}^2 \frac{\exp(\boldsymbol{\theta}_a^\intercal \boldsymbol{x})}{(1+\exp(\boldsymbol{\theta}_a^\intercal \boldsymbol{x}))^2} \right)^{-1}$        Laplace Approximation
return $(\boldsymbol{\mu}'_a, \boldsymbol{\sigma}'_a) \forall a \in \mathcal{A}$
```

**Algorithm 3** Greedy-$K$ Arm Selection with Thompson Sampling

```
Input: Linear models with Gaussian posteriors $(\boldsymbol{\mu}_a, \boldsymbol{\sigma}_a) \forall a \in \mathcal{A}$, number of slots $K$, observed context $\boldsymbol{x}$
Output: $K$ selected arms $a_k$
for $a \in \mathcal{A}$ do
    $\widehat{\mathsf{P}}(R = 1 | X = x; A = a) = \sigma(\boldsymbol{\theta}_a^\intercal \boldsymbol{x})$, where $\boldsymbol{\theta}_a \sim \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\sigma}_a)$ and $\sigma(\cdot)$ denotes the logistic sigmoid
$\pi_g(A|x) = \arg\max_{\pi \in \Omega} \mathop{\mathbb{E}}_{\mathcal{A}_K \sim \pi(A|x)} [\widehat{\mathsf{P}}(C = 1 | X = x; A = \mathcal{A}_K)]$
return $(a_1, \ldots, a_K) \sim \pi_g(A|x)$
```

Laplace approximation then provides an estimate of the updated variance for every model parameter [36, §7.7]. Recent work has shown that this method significantly outperforms several alternatives in a music recommendation set-up, highlighting its competitiveness as a state-of-the-art approach with respect to cumulative reward [3].[1] Traditionally, the focus of competing methods lies on the modelling approach that is used to obtain the estimated probabilities $\widehat{\mathsf{P}}(C = 1 | X = x; A = a)$. The arm selection procedure is then handled as a trivial afterthought: if we want to maximise the expected number of clicks, we simply sort actions by decreasing estimated probability of generating a click and present this top-$K$ list to the user, following the well-known Probability Ranking Principle (PRP) [41]. This greedy arm selection procedure is formalised in Algorithm 3. More involved methods that optimise the entire slate to show to the user exist [20], but their advantages mainly come from increased recommendation utility by taking the interplay of actions into account instead of focusing on fairness of exposure *within* the slate.

### 3.2 Equity of Exposure in Top-$K$ Contextual Bandits

When we show top-$K$ recommendation lists to users, the items ranked at higher positions will have a higher probability of being *examined* by the user than those that are ranked lower. Naturally, recommendations can only be clicked upon when they are examined by the user, and items at the bottom of the ranked list will receive a lower number of clicks regardless of their relevance. This phenomenon is known as *position bias*, and several methods have been proposed in the literature to allow for unbiased learning and evaluation in its presence [25]. We can factorise the process of an item receiving a click (random variable $C$) by (1) the item being exposed to the user (r.v. $E$), and (2) the user perceiving the

[1] Bendada et al. reported that a segmentation-based pseudo-personalisation approach outperformed the Bayesian logistic regression procedure, we found this difference to be attributable to an insufficiently tuned variance hyper-parameter on the Gaussian priors [3].

item as relevant (r.v. $R$):

$$\mathsf{P}(C = 1|X = x; A = a) = \mathsf{P}(R = 1|X = x; A = a)\mathsf{P}(E = 1|X = x; A = a). \tag{2}$$

The relationships between exposure probabilities at different positions are defined by user models [13], encoded by metrics such as Rank-Biased Precision (RBP) [34] or Expected Reciprocal Rank (ERR) [9]. We keep our discussion general for now, but will adopt the *cascading* user model for our case study in Section 3.4.

When deciding which items to show to the user, we are manipulating the exposure probabilities $\mathsf{P}(E)$ in such a way as to maximise the expected exposure for items that are relevant, as this strategy indeed coincides with maximising the expected number of clicks. This makes sense from a (myopic) business perspective, if we suppose that the number of clicks (or more broadly, some notion of positive reward) is indeed the only thing that we need to optimise. Nevertheless, the greedy approach entirely disregards the expected exposure for competing items [16, 44, 49].

In many recommendation scenarios, items are more than their abstract representations would suggest. Multi-sided marketplaces, for example, make up a large portion of recommendation systems on the web [1]. Such platforms connect content *consumers* to content *providers*, with the platform running the recommendation algorithms acting as a broker between the two. Widespread use-cases for such platforms include music streaming, food delivery, e-commerce and employment recommenders. The impact of a recommendation cannot just be measured by a click in these settings, as clicks can significantly influence the financial health of the businesses and people that go behind these "items". Naturally, platforms have limited impact concerning how likely a user is to *like* certain content, but a moral and ethical case can be made to take their impact on *exposure* into account, and *both* are needed to obtain clicks as per Equation 2. Also from a business perspective, the long-term "health" of providers on the platform is crucial to take into account.

Recently, Diaz et al. formalised the principle of equitable expected exposure [16], which informally states that items of equal relevance should receive equal exposure. The fact that item relevance can be dependent on context is what sets this aside from the seminal "equity of attention" principle [5]. Their work focuses on classical information retrieval use-cases with binary or graded relevance labels. These labels can then be used to identify items of equal relevance when distributing exposure. In our use-case, however, we view the relevance of an item conditional on context as a random variable instead of a label. When we consider the (estimated) quantity $\mathsf{P}(R = 1|X = x; A = a)$ to be a graded relevance label, this implies that ranking items greedily (following the PRP and Algorithm 3) optimally satisfies the equity of expected exposure principle. This would be true for any situation where a total order is defined over the set of items. Nevertheless, this does not account for the fact that differences in relevance might be arbitrarily small ($0 \leq \mathsf{P}(R = 1|A = a) - \mathsf{P}(R = 1|A = a') < \epsilon$), whereas the difference in exposure might be significant (e.g. in a top-1 list, $\mathsf{P}(E = 1|A = a) = 1, \mathsf{P}(E = 1|A = a') = 0$ would be seen as "optimal"). Making the trade-off explicit, Diaz et al. decompose their proposed objective into "Expected Exposure Disparity" (EE-D) and "Expected Exposure Relevance" (EE-R), proposing to optimise the area under the disparity-relevance curve (EE-AUC) [16]. Their notion of EE-R, w.r.t. the "optimal" greedy ranking, does not take into account the magnitude of the difference between relevance probabilities for competing arms. Furthermore, it is non-trivial to compute the impact of a specific top-$K$ selection of arms on the EE-AUC metric, and it has not been studied for bandit use-cases that deal with online, delayed updates.

Instead, we wish to retain the focus of the contextual bandit paradigm on that of maximising expected reward. To model the trade-off between the desired equity of exposure and the number of clicks, we introduce a tolerance parameter $\epsilon$. The goal of an exposure-aware arm selection procedure is then to minimise the disparity in the exposure distribution whilst losing at most $\epsilon$ expected clicks. This trade-off between relevance and fairness is often seen as a phenomenon that occurs over the global user-base, and a global hyper-parameter is then introduced to decide just how important equity of exposure should be without impacting the user-focused utility metrics "too much". Recent work studies how different users respond to *diverging* recommendations, and observed that the *openness* of a user to randomised recommendations can vary wildly [32]. Furthermore, this property of a user can be predicted based on behavioural data on the platform. This finding hints that we can tackle the trade-off in a more effective manner via targeted randomisation to the users that allow it, instead of blindly doing this for the entire user-base, possibly hurting the user experience for more *focused* users in the meanwhile. We illustrate such a hypothetical setting in Figure 1, where the $K^{\text{th}}$ best recommendation might still be a good choice for some users (*diverse* and *average* users), but worse for others (*focused* users). Figure 1 shows a toy example that serves to illustrate – but was inspired by real-world data.

### 3.3 Personalising the Relevance-Fairness Trade-off with Targeted Randomisation

The x-axis in Figure 1 shows the top-12 items for a given user, the y-axis shows the probability of relevance over these items for that user. For the *focused* user, the probability decreases rather quickly after the top-1 item. This indicates that there is little room for randomisation, because ranking the top item further down the list forces us to fill the preceding spots with clearly sub-optimal items. This, in turn, can hurt both the user-experience and typical business metrics that are measured through the platform. In contrast, the *diverse* user demonstrates a much more favourable relevance-fairness trade-off, as the probability of relevance for the $5^{\text{th}}$-best item is about 90% of that of the best item. We can observe two consequences for such diverse users: (1) there will be high exposure disparity under the greedy-$K$ approach, as the exposure at positions 1 and 5 will not be proportional to their difference in relevance, and (2) this disparity in exposure can be alleviated with a minimal loss in expected utility.

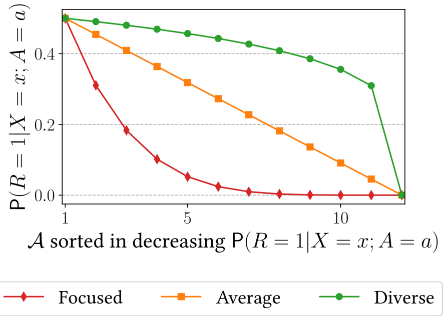


Fig. 1. Top-$K$ relevances for varying types of users.

Before we move on, we must formally define what we consider to be disparate exposure. Following the equity of exposure principle, we posit that the probability of a click on an item should be proportional to the probability of relevance for that item. That is, the distribution for $\mathsf{P}(A|C)$ (which we can control) should match that of $\mathsf{P}(A|R)$ (which is fixed and never known exactly). An algorithm that achieves fairness of exposure can then be seen as one that minimises the statistical divergence between these two conditional probability distributions, denoted as $\mathsf{D}(\mathsf{P}(A|R); \mathsf{P}(A|C))$. In practice, we can use the model estimates for $\widehat{\mathsf{P}}(R)$ in concordance with a user model to provide estimates for $\widehat{\mathsf{P}}(E)$ and the resulting distribution of $\widehat{\mathsf{P}}(C)$. Algorithm 4 formalises the Exposure-Aware aRm Selection (EARS) algorithm that minimises $\mathsf{D}(\mathsf{P}(A|R); \mathsf{P}(A|C))$ whilst losing at most $\epsilon$ relative expected clicks in comparison with the Greedy-$K$ procedure presented in Algorithm 3. As such, we are maximising fairness whilst remaining within a configurable acceptable distance from the performance of the optimal policy.

Various $f$-divergences can be used in a concrete instantiation of our method, we will focus on the widely used Hellinger distance in our experiments (denoted by $\mathsf{D}_{\mathsf{H}}$) [29]. Most divergences will treat the most and least relevant items with the same weight, giving rise to spurious behaviour from the long tail. To mitigate this, we focus on the Hellinger distance between the relevance- and click-distributions for the top-$K$ items only. Note that without the tolerance parameter $\epsilon$, minimising the divergence between the relevance- and click-distributions is straightforward. Clicks depend on exposure and relevance, per definition. If we want clicks to *only* depend on the relevance $R$, we need

**Algorithm 4** Exposure-Aware aRm Selection (EARS) with Thompson Sampling

```
Input: Linear models with Gaussian posteriors $(\boldsymbol{\mu}_a, \boldsymbol{\sigma}_a) \forall a \in \mathcal{A}$, number of slots $K$, observed context $\boldsymbol{x}$, tolerance $\epsilon$
Output: $K$ selected arms $a_k$
for $a \in \mathcal{A}$ do
    $\widehat{\mathsf{P}}(R = 1|X = x; A = a) = \sigma(\boldsymbol{\theta}_a^{\intercal}\boldsymbol{x})$, where $\boldsymbol{\theta}_a \sim \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\sigma}_a)$ and $\sigma(\cdot)$ denotes the logistic sigmoid
$\pi_g(A|x) = \underset{\pi\in\Omega}{\arg\max} \underset{\mathcal{A}_K\sim\pi(A|x)}{\mathbb{E}} [\widehat{\mathsf{P}}(C = 1|X = x; A = \mathcal{A}_K)]$
$\pi_e(A|x) = \underset{\pi\in\Omega}{\arg\min} \underset{\mathcal{A}_K\sim\pi(A|x)}{\mathbb{E}} \mathrm{D}\left(\widehat{\mathsf{P}}(A|R); \widehat{\mathsf{P}}(A|C)\right)$ subject to $\left(1 - \frac{\mathbb{E}_{\pi_e}[C]}{\mathbb{E}_{\pi_g}[C]}\right) \leq \epsilon$
return $(a_1, \ldots, a_K) \sim \pi_e(A|x)$
```

to remove the causal link between $E$ and $C$ [39]. This can be achieved via randomisation, which effectively boils down to shuffling the item catalogue instead of ranking items according to their estimated relevance probabilities.

When we do have a tolerance parameter $\epsilon$, we need to find those items that imply the largest decrease in disparity for the smallest decrease in clicks. The largest decrease in disparity can be achieved by randomly shuffling items. In particular, the divergence will be highest among the most relevant items, either for those receiving proportionally too many or not enough clicks. Because of this, we should focus our efforts on the head rather than the tail of the catalogue. Naturally, re-distributing exposure from the 1st to the 2nd most relevant item implies the lowest possible decrease in clicks as well (compared to any other target item), reinforcing the observation that we should focus on the most relevant items. As a consequence, optimising the exposure-aware arm selection strategy corresponds to *shuffling* the top-$K'$ items, instead of *sorting* them by their estimated relevance probabilities, whilst sorting the rest of the list. As such, the problem is reduced to finding the maximal value of $K'$ that implies a reward decrease of at most $\epsilon$. Computing an expectation over clicks requires us to adopt a user model that describes the probability of exposure for every item in a given ranking. The methods we propose in this manuscript are generally applicable and not tied to this choice. In what follows, we discuss an instantiation of the top-$K$ contextual bandit problem, and discuss how our exposure-aware arm selection strategy may be applied efficiently and effectively.

### 3.4 Case Study: Equity of Exposure in Carousel-Based Music Recommendation

To make matters concrete, we focus on the use-case of carousel personalisation for music recommendation. As real-world data is publicly available for this application, it allows for reproducible evaluation of our proposed method [3]. The work of Bendada et al. presents this use-case as a contextual bandit problem with multiple plays, providing implementations of several competing methods based on $\epsilon-$greedy, UCB and Thompson Sampling. Their focus lies on recommendation *carousels*, which are widely used to present personalised content to users in music streaming apps.

In a recommendation carousel, recommendations that are ranked lower in the list only become visible to the user after they have explicitly scrolled the carousel past earlier items. Additionally, the assumption is made that users will *leave* the carousel once they have found relevant content. As such, when a user clicks on an item at position $i$, the items at positions $i < j \leq K$ are no longer exposed to the user. This corresponds to the well-know *cascade* user model [10], which is also adopted by the widely used ERR metric [9]. We adopt the generalisation of this that includes an *abandonment* probability, explicitly modelling that users might choose to stop examining the carousel at every position when they still haven't found what they were looking for. Indeed, without this abandonment parameter, the expected number of clicks on a carousel is only dependent on the set of included items and not on the ordering within the carousel. This leads to spurious behaviour, where sorting the carousel by *ascending* relevance in the top-$K$ maximises exposure (as the user is more likely to keep scrolling) without having any effect on the total expected reward.

Let $\sigma = (\sigma_1, \ldots, \sigma_K)$ denote a ranking of $K$ actions, obtained from the recommendation policy $\pi$. When $(1 - \gamma)$ denotes the abandonment probability, we define the exposure probability for the item at position $k$ as the probability that *all* higher-ranked items were *not* found relevant and the user has not abandoned the carousel:

$$\mathsf{P}(E = 1 | \text{pos} = k) = \gamma^k \prod_{1 \leq j < k} (1 - \mathsf{P}(R = 1 | A = \sigma_j)). \tag{3}$$

Bringing Equation 3 together with Equation 2 to compute the probability of a click on a given recommendation in a carousel filled according to the ranking $\sigma$, we obtain:

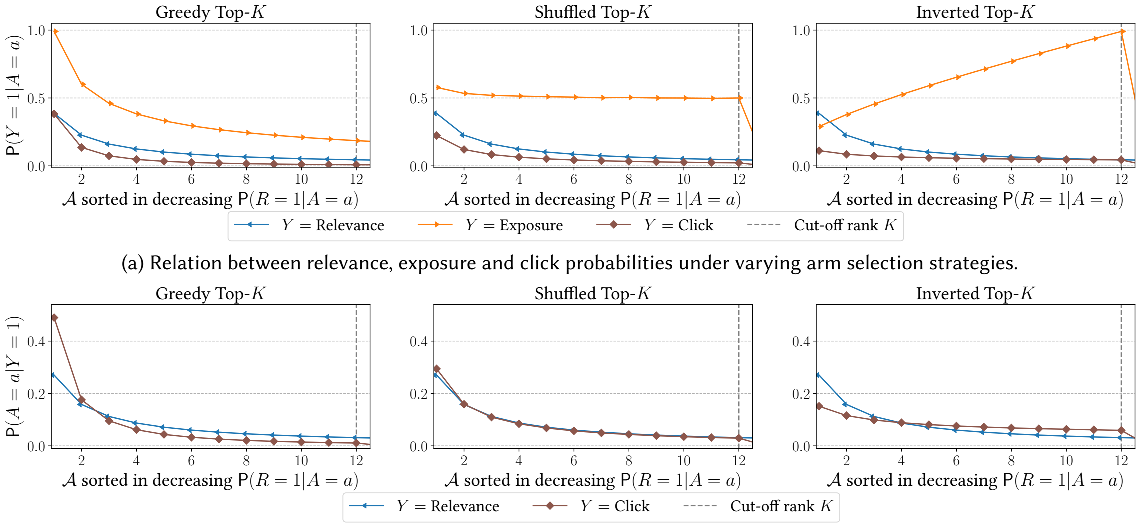
$$\begin{aligned} \mathsf{P}(C = 1 | X = x; A = \sigma_k; \text{pos} = k) &= \mathsf{P}(R = 1 | X = x; A = \sigma_k) \cdot \mathsf{P}(E = 1 | X = x; \text{pos} = k) \\ &= \mathsf{P}(R = 1 | X = x; A = \sigma_k) \gamma^k \prod_{1 \leq j < k} (1 - \mathsf{P}(R = 1 | X = x; A = \sigma_j)) \end{aligned} \tag{4}$$

Naturally, the total number of expected clicks is the sum over all positions. When we fill in estimated relevance probabilities obtained via a bandit algorithm (e.g. sampled from approximate posteriors as per Alg. 2), we can compute an estimate of the expected exposure at every position, as well as the expected number of clicks for any given ranking.

From the user-item interaction data provided by Bendada et al. [3], we can obtain the relevance distributions over top-$K$ items for every user (as in Fig. 1). We average the relevance probabilities over the top-12 items for every user in the dataset and show the resulting estimates in Figure 2. Compared to Figure 1, we see that the average user in the dataset is somewhere between what we called "Focused" and "Average" in our example. For this obtained prototype of a user, we can visualise how competing arm selection strategies influence the exposure and click probabilities for every item in the top-$K$ carousel. We compare the greedy arm selection strategy (Alg. 3), the exposure-aware arm selection strategy without a tolerance parameter $\epsilon$ (Alg. 4), and a strategy that inverts the greedy top-$K$ ranking. We set $\gamma = 0.99$.

From Figure 2a, we can see that the top-1 item is almost always exposed to the user in the greedy ranking ($\mathsf{P}(E = 1 | \text{pos} = 1) = \gamma$). This leads to a steep decline in exposure probability at lower positions, which follows the shape of the relevance probabilities. However, as the click probability is the product of exposure and relevance, we can see a much steeper decline for the click probabilities. That is, there are disproportionately more clicks on the top item compared to the lower ranked items, with respect to their relevance. As is expected, we can flatten the exposure curve and provide equal exposure over all items by making the arm selection procedure stochastic as well. When shuffling the entire top-$K$ carousel before every impression, we observe that we can obtain a more equitable allocation of exposure. Note that exposure is not entirely flattened, as the cascade user model still posits that an item ranked after a relevant item has a lower probability of being examined than if it had been ranked after a non-relevant item. Because of this, the most relevant item ranked at position $K$ has a higher exposure probability than the least relevant item at the same position. For high abandonment probabilities (low $\gamma$), this effect is dampened. As we can see for the inverted greedy arm selection strategy, now the top-$K^{\text{th}}$ item has exposure probability $\gamma$, and it declines steadily over the remainder of the carousel.

From these insights, we turn to the question we are trying to answer: "Given a *click* on an item, how likely was it that the item was $i$?" (i.e. $\mathsf{P}(A = i | C = 1)$). The equity of exposure principle states that the answer to this question should be similar to the answer for "Given the user found an item *relevant*, how likely was it that the item was $i$?" (i.e. $\mathsf{P}(A = i | R = 1)$). Figure 2b visualises these conditional probability distributions over the top-$K$ for competing arm



(a) Relation between relevance, exposure and click probabilities under varying arm selection strategies.

(b) Relation between action probabilities conditional on clicks and relevance under varying arm selection strategies.

Fig. 2. Visualising how exposure is distributed over items for varying arm selection strategies.

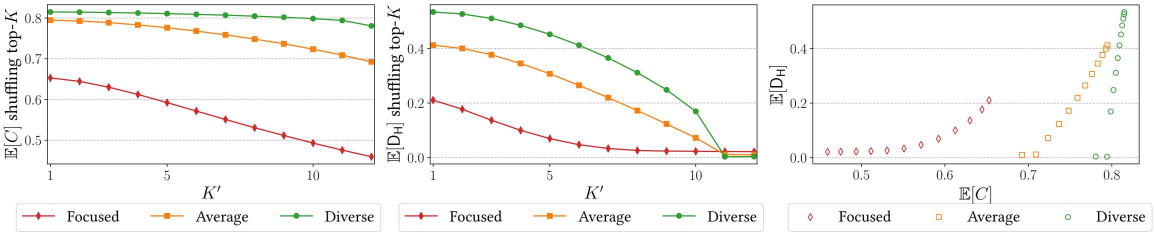
selection strategies. It clearly shows how, for the greedy strategy, the top-2 items are disproportionately likely to receive clicks according to their relevances. The top-3 to top-12 items are under-exposed as a result. The inverted strategy inverts this trend, and the shuffled strategy is able to alleviate virtually all divergence between the two distributions. It should be clear that this arm selection strategy is favourable w.r.t. $\mathsf{D}\left(\mathsf{P}(A|R); \mathsf{P}(A|C)\right)$. When the tolerance parameter $\epsilon$ comes into play, $\mathsf{P}(A = i|C = 1)$ will be slightly lifted for the top-$K'$ and lowered for the top-$K'$-to-$K$, w.r.t. a full reshuffling.

### 3.4.1 *Different Strokes for Different Folks: Personalising the Relevance-Fairness Trade-off.*

We now look into the impact of varying the size of the reshuffled head, and how it impacts expected clicks and disparity for various types of users. Figure 3a shows the expected number of clicks on the top-$K$ for our running example users, for varying shuffle lengths $K'$. We can clearly see that more reshuffling is allowed for the diverse and average users, compared to the focused user. Figure 3b visualises the expected disparity under these arm selection strategies (i.e. the Hellinger distance between $\mathsf{P}(A|R)$ and $\mathsf{P}(A|C)$). It is clear that reshuffling more head items decreases disparity, that disparity was higher for the more diverse users to begin with, and that most disparity can be alleviated at low values of $K$ for the focused user. A natural trade-off between maximising clicks and minimising disparity arises, which we visualise in Figure 3c. We can gain a lot in equitability of exposure by reshuffling the entire top-$K$ for the diverse user, with a minimal cost in expected clicks. For the focused user, however, we can significantly decrease disparity at lower values of $K'$, but the cost in terms of expected clicks is higher. Based on estimates for $\widehat{\mathsf{P}}(R|X;A)$, we can handle this trade-off in real-time.

### 3.4.2 *Efficiently Computing Expected Clicks for the Cascade Model.*

An efficient implementation of Algorithm 4 requires a method to efficiently compute the expected clicks under the greedy arm selection strategy, as well as under the shuffled strategy. As the greedy approach is deterministic, this boils down to:



(a) Expected clicks. (b) Expected disparity. (c) Click-Disparity Trade-off

Fig. 3. Different types of users impact the expected number of clicks and the expected disparity in item exposure in different ways under our proposed exposure-aware arm selection procedure. Diverse users allow for the disparity to be significantly reduced with only a marginal impact on the expected number of clicks. Focused users manifest less disparity to begin with, and reshuffling with large $K$ can significantly decrease expected clicks.

$$\begin{aligned}\mathop{\mathbb{E}}_{\sigma\sim\pi_g}[C|X=x] &= \sum_{1\leq i\leq K} \mathrm{P}(C=1|X=x;A=\sigma_i;\mathrm{pos}=i)\\ &= \sum_{1\leq i\leq K} \mathrm{P}(R=1|X=x;A=\sigma_i)\mathrm{P}(E=1|X=x;A=\sigma_i;\mathrm{pos}=i).\\ &= \sum_{1\leq i\leq K} \mathrm{P}(R=1|X=x;A=\sigma_i)\gamma^i \prod_{1\leq j<i}(1-\mathrm{P}(R=1|X=x;A=\sigma_j)).\end{aligned} \tag{5}$$

As the reshuffling operation for the exposure-aware arm selection procedure is stochastic, computing an expectation for clikcs under $\pi_e$ requires us to consider all possible permutations of the top-$K$ list. The symmetric group over the set of items – that is, the set containing all permutations $\sigma$ that correspond to rankings of the top-$K$ items – is defined as $S_K$. When we shuffle the top-$K$ ($\pi_e$), all permutations are equally likely: $\mathrm{P}(\sigma)=\frac{1}{K!}$. We can compute the expected number of clicks by averaging over all possible permutations:

$$\mathop{\mathbb{E}}_{\sigma\sim\pi_e}[C|X=x] = \sum_{\sigma\in S_K}\mathrm{P}(\sigma)\sum_{1\leq i\leq K}\mathrm{P}(R=1|X=x;A=\sigma_i)\gamma^i\prod_{1\leq j<i}(1-\mathrm{P}(R=1|X=x;A=\sigma_j)). \tag{6}$$

Because this is a combinatorial problem, computing the expected number of clicks under random shuffling by explicitly listing all possible permutations quickly becomes intractable as $K$ grows. We can, however, see that many possible rankings are considered equivalent up to a certain point in Equation 6. Indeed, the product over the items at ranks $j\in[1,i)$ does not depend on the ordering of items at said ranks, leading to $(i-1)!$ equivalent permutations. We can explicitly take these equivalent permutations into account and scale down the combinatorial space of permutations to the exponential space of the powerset. Intuitively, for every possible item $\sigma_i$ we look at every possible subset of other items $\sigma_j$ that could have been ranked higher, as the probability of exposure for $\sigma_i$ is dependent on that set. For every such subset $\mathcal{J}\in 2^{\sigma\setminus\sigma_i}$, there are $|\mathcal{J}|!$ possible permutations of higher ranked items, and $(K-|\mathcal{J}|-1)!$ possible permutations for lower ranked items to which the click probability is indifferent:

$$\mathop{\mathbb{E}}_{\sigma\sim\pi_e}[C|X=x] = \sum_{1\leq i\leq K}\mathrm{P}(R=1|X=x;A=\sigma_i)\gamma\sum_{\mathcal{J}\in 2^{\sigma\setminus\sigma_i}}\gamma^{|\mathcal{J}|}\frac{(|\mathcal{J}|)!(K-|\mathcal{J}|-1)!}{K!}\prod_{j\in\mathcal{J}}(1-\mathrm{P}(R=1|X=x;A=j))\,. \tag{7}$$

Computing this formula straightforwardly leads to $O(K2^{K-1}(K-1))$ computations, which is already a significant improvement from the original combinatorial complexity of $O(K!)$. We can recognise the normalising factor that defines the fraction of permutations that correspond to the same exposure probability as the Beta function: $B(|\mathcal{J}|+1, K-|\mathcal{J}|)$. To further simplify the computation, we take the inverse probability $\mathsf{P}(C=1) = 1 - \mathsf{P}(C=0)$. If we define $2^{\sigma}_{k} = \{\mathcal{J} \in 2^{\sigma} \mid |\mathcal{J}| = k\}$ as all subsets of length $k$ from the top-$K$ (denoted by $\sigma$), then:

$$\mathop{\mathbb{E}}_{\sigma\sim\pi_e}[C|X=x] = 1 - \left(\sum_{k=0}^{K} \frac{\gamma^{k}(1-\gamma)^{\mathbb{1}\{K=k\}}}{\binom{K}{k}} \sum_{\mathcal{J}\in 2^{\sigma}_{k}} \prod_{j\in\mathcal{J}} \mathsf{P}(R=0|X=x;A=j)\right). \tag{8}$$

When reshuffling only the first $K'$ items instead of the full top-$K$, this decomposes as follows:

$$\begin{aligned}\mathop{\mathbb{E}}_{\sigma\sim\pi_e}[C|X=x] = 1-&\left(\left(\sum_{k=0}^{K'} \frac{\gamma^{k}(1-\gamma)^{\mathbb{1}\{K=k\}}}{\binom{K'}{k}} \sum_{\mathcal{J}\in 2_{k}^{\sigma^{\star}_{1:K'}}} \prod_{j\in\mathcal{J}} \mathsf{P}(R=0|X=x;A=j)\right)\right.\\ &\left.+\left(\sum_{k=K'}^{K} \gamma^{k}(1-\gamma)^{\mathbb{1}\{K=k\}} \prod_{j\in\sigma^{\star}_{1:k}} \mathsf{P}(R=0|X=x;A=j)\right)\right).\end{aligned} \tag{9}$$

Here, $\sigma^{\star}_{1:K'}$ denotes the top-$K'$ items according to the optimal greedy ranking. Intuitively, the first term needs to account for the permutations of the first $K'$ items that are being reshuffled, whereas the second term handles the remaining $(K-K')$ items that always have a fixed position in the carousel. To compute the maximal reshuffling size $K'$ that minimises disparity within the tolerance $\epsilon$, we compute the expected clicks for varying values of $K'$ using Equation 9 and act accordingly. Because this formula can be computed efficiently (in the order of milliseconds for $K=12$), we can perform these computations in real-time during the arm selection process.

## 4 EXPERIMENTAL RESULTS AND DISCUSSION

To validate the effectiveness of our proposed algorithm, we implement EARS in the carousel personalisation simulation framework presented by Bendada et al. [3]. The framework consists of real-world data from the Deezer music streaming platform, with relevance information between 862 playlists and 974 960 users. Users and items are both represented by a dense 97-dimensional feature vector. User features are presented as input to competing bandit algorithms ($\boldsymbol{x}$ in Algorithm 4), whereas item features remain hidden and are only used to compute the ground-truth relevances scores between user-item pairs. Every *round*, users are randomly sampled with replacement and presented to the bandit algorithm, which yields a top-$K$ list of recommendations to present to the user. Following the framework, we set $K=12$. Rewards are then simulated following the cascade user model, and the contextual bandit is updated in batch every 20 000 users ($T_r = 20\,000$). We refer the interested reader to the work of Bendada et al. or their open-source repository for more information on the experimental setup.[2] Originally, their simulator supports the cascade user-model *without* abandonment parameter $\gamma$. As we have laid out above, this assumption makes the relevance-fairness trade-off trivial: exposure can be maximised by ranking less relevant items first, as the user is then more likely to keep scrolling through the carousel. We set $\gamma$ to 0.9 in our experiments, implying that the exposure probability for the item at position 12 is discounted by a factor of $0.9^{12} = 0.282$ regardless of the relevance of the list.

Bendada et al.'s simulation environment contains implementations for various (pseudo-)personalisation algorithms based on the multi-armed and contextual bandit paradigms. They report that a segmentation-based Thompson Sampling

[2] https://github.com/deezer/carousel_bandits

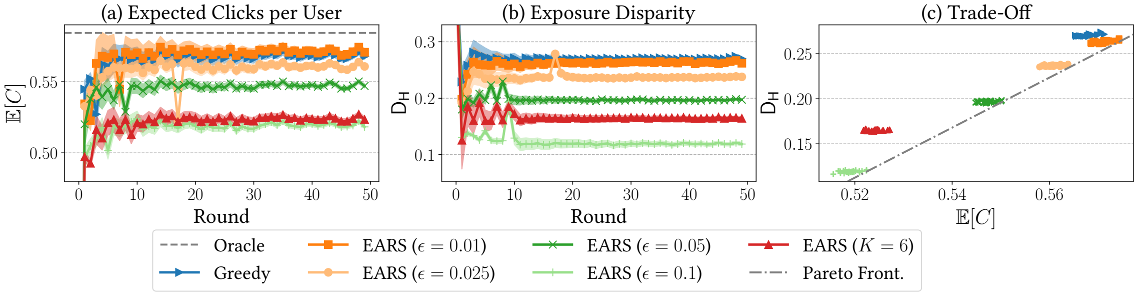


Fig. 4. Experimental results for our EARS algorithm on top of Logistic Regression Thompson Sampling compared to the greedy ranking approach and the optimal, unattainable oracle. We include EARS with a tolerance parameter $\epsilon$ that personalises the shuffling parameter and EARS with a fixed value of $K'$. We observe that EARS with a personalised shuffling parameter moves the Pareto frontier favourably, indicating an improved relevance-fairness trade-off.

approach that follows a Beta-Bernoulli model outperforms all competing approaches, including the online Bayesian logistic regression algorithm presented in Algorithm 2 [8]. We noticed that the reason for this was that the variance hyper-parameter on the priors was insufficiently tuned, and set it to 2 000 in our experiments instead of the original unit variance (this parameter was found via grid search). After this change, the linear Thompson Sampling method significantly and consistently improves over all alternatives, converging quickly to near-optimal parameters. For these reasons, we consider the effectiveness of EARS on top of this state-of-the-art approach. All source code to reproduce our experimental results is available at github.com/olivierjeunen/EARS-recsys-2021. The research questions we wish to answer with out experimental setup, are the following:

**RQ1** How does EARS impact the expected reward per round for a state-of-the-art bandit algorithm?
**RQ2** Is EARS effective in minimising disparity of exposure?
**RQ3** Does EARS maintain a favourable relevance-fairness trade-off?

***RQ1**: EARS' Impact on Expected Reward.* Figure 4a shows the expected reward per user at every round for competing algorithms, with a shaded 95% confidence interval. We compare both the greedy ranking approach, personalised EARS with $\epsilon \in \{0.01, 0.025, 0.05, 0.1\}$, and a non-personalised EARS variant that shuffles the top-6 for every user. Additionally, we report the expected number of clicks obtained by the optimal oracle policy that greedily ranks items according to their true relevance probabilities. We observe that all variants start to converge after a few rounds, and are fully converged after roughly 15 to 20 rounds. The gap in expected reward between the oracle and the greedy approach (i.e. regret) is around 0.013, which is far more favourable than any other competing bandit algorithm implemented in the framework (based on UCB, $\epsilon$-greedy exploration or pseudo-personalisation). When increasing EARS' tolerance parameter $\epsilon$, the expected reward starts to decrease gracefully as theory would suggest. With a fixed reshuffling size of $K = 6$, the loss in expected reward is comparable to that when $\epsilon = 0.1$. Additionally, we observe that the expected reward for EARS with $\epsilon = 0.01$ slightly *increases* over the greedy approach. This can be attributed to two reasons: (1) As our logistic regression model does not correct for position bias, we are effectively learning a model for $\mathsf{P}(C)$ instead of $\mathsf{P}(R)$. EARS partially mitigates such biases by randomising the top-$K$ list, which benefits the model. (2) Randomising the top-$K$ list additionally leads to more diverse training data at every round, also benefiting the model. Finally, note that the $\epsilon$-parameter denotes the tolerance in loss of expected reward with respect to the greedy ranking from the same model, instead of the greedy model shown in the plot. In conjunction with the two reasons laid out above, EARS should not be expected to achieve 99% performance of the greedy approach that is shown, as these models tackled the exploration-exploitation trade-off slightly differently as well. If we stop the randomisation process for EARS after the model parameters are converged, we observe an improvement in expected reward with magnitude $\epsilon$.

***RQ2**: EARS' Impact on Exposure Disparity.* Figure 4b shows the average disparity in exposure over the top-12 items at every round for competing algorithms, with a shaded 95% confidence interval. We observe that, comparable to expected reward, all competing algorithms tend to converge rather quickly and exhibit stable behaviour afterwards. As expected, we observe that the greedy approach attains maximal disparity, and that this disparity is gradually decreased for EARS variants as $\epsilon$ increases. Although EARS with a fixed $K' = 6$ obtains a comparable expected reward as EARS with $\epsilon = 0.1$ (as per Figure 4a), we observe that its disparity in exposure is significantly larger. This reassures us that peronalising the trade-off pays off. For focused users, a fixed $K'$ incurs a larger loss in expected reward than can be justified by the expected loss in exposure disparity.

***RQ3**: EARS' Relevance-Fairness Trade-Off.* Figure 4c shows the trade-off between expected reward (x-axis) and fairness of exposure (y-axis) for competing approaches. We plot all measurements from Figures 4a and 4b from rounds 20 and onward, to compare the performance of the methods when model parameters have converged. The best approaches are those reaching the lower right quadrant of the plot, as they are effective in maximising reward whilst minimising disparity. First, we observe that EARS with $\epsilon = 0.01$ manages to both improve on expected reward and decrease disparity compared to the greedy approach. As explained in the paragraph detailing **RQ1** above, this can be attributed to the favourable effects of randomisation on the informational value of the gathered training data. Second, we can now clearly see the linear nature of the trade-off between expected clicks and expected disparity. The line connecting measurements for varying values of $\epsilon$ denotes the Pareto-frontier for EARS: methods that lead to measurements above this line are Pareto-dominated by EARS, as an improvement in either fairness or expected reward can then be achieved without hurting the other objective. This is the case for non-personalised EARS, which can benefit from significantly improved fairness without hurting expected reward by personalising the trade-off. Varying the tolerance parameter $\epsilon$ allows us to move along the Pareto-frontier, in favour of either fairness or reward. Note that this is not specific to the adopted Hellinger distance. We have looked into other measures of divergence between probability distributions such as Kullback-Leibler divergence and Total Variation, and consistently observed the same phenomenon.

## 5 CONCLUSION AND FUTURE WORK

Recommendation platforms that are deployed in the real world affect the world around them. In marketplaces and multi-stakeholder scenarios, unfair distributions of attention or exposure can propagate to unfair distributions of income for content providers. In this work, we have studied how such problems of disparate exposure relate to top-$K$ contextual bandit recommendation problems. We have formalised this problem setting in a probabilistic mathematical framework, and have proposed a general Exposure-Aware aRm Selection (EARS) algorithm that tackles the fairness-relevance trade-off in a personalised manner. We have presented an application of our newly proposed method to carousel personalisation for music recommendation, both showing how it effectively minimises disparate exposure whilst minimally impacting expected reward, and how it can be efficiently applied to calculate the expected loss in reward in real-time for cascade user models. Experimental results from publicly available, real-world music streaming data indicate that EARS is effective in decreasing disparate exposure whilst remaining effective in terms of expected reward. EARS decides how to deal with arm selection based on relevance estimates, independent of how those relevance estimates were obtained. As a result, it can easily be applied on top of any existing bandit system with minimal changes.

In future work, we wish to extend our analysis to larger action spaces and broader applications with other user models than the cascade model. Additionally, we wish to study how disparate exposure occurs in reinforcement learning recommendation applications, instead of the simplified bandit model that is more commonly studied in the literature, and how our findings translate to off-policy settings [22]. Finally, it would be interesting to expand EARS to ensure group-based equity of exposure and other notions of fairness in recommendation scenarios.

## ACKNOWLEDGMENTS

This work received funding from the Flemish Government (AI Research Program). We thank Joey De Pauw for insightful discussions and feedback.

## REFERENCES

[1] H. Abdollahpouri, G. Adomavicius, R. Burke, I. Guy, D. Jannach, T. Kamishima, J. Krasnodebski, and L. Pizzato. 2020. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction* 30, 1 (2020), 127–158.

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

[3] W. Bendada, G. Salha, and T. Bontempelli. 2020. Carousel Personalization in Music Streaming Apps with Contextual Bandits. In *Proc. of the 14th ACM Conference on Recommender Systems (RecSys '20)*. ACM, 420–425.

[4] A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi, and C. Goodrow. 2019. Fairness in Recommendation Ranking through Pairwise Comparisons. In *Proc. of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '19)*. ACM, 2212–2220.

[5] A. J. Biega, K. P. Gummadi, and G. Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *Proc. of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '18)*. ACM, 405–414.

[6] A. Bower, H. Eftekhari, M. Yurochkin, and Y. Sun. 2021. Individually Fair Rankings. In *International Conference on Learning Representations (ICLR '21)*.

[7] R. Burke. 2017. Multisided Fairness for Recommendation. *CoRR* abs/1707.00093 (2017). arXiv:1707.00093

[8] O. Chapelle and L. Li. 2011. An Empirical Evaluation of Thompson Sampling. In *Proc. of the 24th International Conference on Neural Information Processing Systems (NIPS'11)*. 2249–2257.

[9] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proc. of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)*. ACM, 621–630.

[10] O. Chapelle and Y. Zhang. 2009. A Dynamic Bayesian Network Click Model for Web Search Ranking. In *Proc. of the 18th International Conference on World Wide Web (WWW '09)*. ACM, 1–10.

[11] M. Chen, A. Beutel, P. Covington, S. Jain, F. Belletti, and E. H. Chi. 2019. Top-K Off-Policy Correction for a REINFORCE Recommender System. In *Proc. of the 12th ACM International Conference on Web Search and Data Mining (WSDM '19)*. ACM, 456–464.

[12] Y. Chen, A. Cuellar, H. Luo, J. Modi, H. Nemlekar, and S. Nikolaidis. 2020. Fair Contextual Multi-Armed Bandits: Theory and Experiments. In *Proc. of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, Vol. 124. PMLR, 181–190.

[13] A. Chuklin, I. Markov, and M. de Rijke. 2015. Click models for web search. *Synthesis lectures on information concepts, retrieval, and services* 7, 3 (2015), 1–115.

[14] M. F. Dacrema, P. Cremonesi, and D. Jannach. 2019. Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches. In *Proc. of the 13th ACM Conference on Recommender Systems (RecSys '19)*. ACM, 101–109.

[15] Y. Deldjoo, V. W. Anelli, H. Zamani, A. Bellogín, and T. Di Noia. 2021. A flexible framework for evaluating user and item fairness in recommender systems. *User Modeling and User-Adapted Interaction* (2021).

[16] F. Diaz, B. Mitra, M. D. Ekstrand, A. J. Biega, and B. Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure. In *Proc. of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*. ACM, 275–284.

[17] B. Dumitrascu, K. Feng, and B. E. Engelhardt. 2018. PG-TS: Improved Thompson Sampling for Logistic Contextual Bandits. In *Advances in Neural Information Processing Systems 30 (NeurIPS'18)*. 4629–4638.

[18] F. Garcin, B. Faltings, O. Donatsch, A. Alazzawi, C. Bruttin, and A. Huber. 2014. Offline and Online Evaluation of News Recommender Systems at Swissinfo.Ch. In *Proc. of the 8th ACM Conference on Recommender Systems (RecSys '14)*. 169–176.

[19] A. Gilotte, C. Calauzènes, T. Nedelec, A. Abraham, and S. Dollé. 2018. Offline A/B Testing for Recommender Systems. In *Proc. of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. ACM, 198–206.

[20] E. Ie, V. Jain, J. Wang, S. Narvekar, R. Agarwal, R. Wu, H. Cheng, T. Chandra, and C. Boutilier. 2019. SlateQ: A Tractable Decomposition for Reinforcement Learning with Recommendation Sets. In *Proc. of the 28th International Joint Conference on Artificial Intelligence (IJCAI '19)*. 2592–2599.

[21] O. Jeunen. 2019. Revisiting Offline Evaluation for Implicit-Feedback Recommender Systems. In *Proc. of the 13th ACM Conference on Recommender Systems (RecSys '19)*. ACM, 596–600.

[22] O. Jeunen and B. Goethals. 2021. Pessimistic Reward Models for Off-Policy Learning in Recommendation. In *Proc. of the 15th ACM Conference on Recommender Systems (RecSys '21)*. ACM.

[23] O. Jeunen, D. Rohde, and F. Vasile. 2019. On the Value of Bandit Feedback for Offline Recommender System Evaluation. In *Proc. of the ACM RecSys Workshop on Reinforcement Learning and Robust Estimators for Recommendation (REVEAL '19)*.

[24] O. Jeunen, D. Rohde, F. Vasile, and M. Bompaire. 2020. Joint Policy-Value Learning for Recommendation. In *Proc. of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '20)*. ACM, 1223–1233.

[25] T. Joachims, A. Swaminathan, and T. Schnabel. 2017. Unbiased Learning-to-Rank with Biased Feedback. In *Proc. of the 10th ACM International Conference on Web Search and Data Mining (WSDM '17)*. ACM, 781–789.

[26] M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth. 2016. Fairness in Learning: Classic and Contextual Bandits. In *Advances in Neural Information Processing Systems (NeurIPS '16)*, Vol. 29.

[27] C. Li, H. Feng, and M. de Rijke. 2020. Cascading Hybrid Bandits: Online Learning to Rank for Relevance and Diversity. In *Proc. of the 14th ACM Conference on Recommender Systems (RecSys '20)*. ACM, 33–42.

[28] L. Li, W. Chu, J. Langford, and R. E. Schapire. 2010. A Contextual-Bandit Approach to Personalized News Article Recommendation. In *Proc. of the 19th International Conference on World Wide Web (WWW '10)*. ACM, 661–670.

[29] F. Liese and I. Vajda. 2006. On Divergences and Informations in Statistics and Information Theory. *IEEE Transactions on Information Theory* 52, 10 (2006), 4394–4412.

[30] R. Mehrotra and B. Carterette. 2019. Recommendations in a Marketplace. In *Proc. of the 13th ACM Conference on Recommender Systems (RecSys '19)*. ACM, 580–581.

[31] R. Mehrotra, J. McInerney, H. Bouchard, M. Lalmas, and F. Diaz. 2018. Towards a Fair Marketplace: Counterfactual Evaluation of the Trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems. In *Proc. of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. ACM, 2243–2251.

[32] R. Mehrotra, C. Shah, and B. Carterette. 2020. Investigating Listeners' Responses to Divergent Recommendations. In *Proc. of the 14th ACM Conference on Recommender Systems (RecSys '20)*. ACM, 692–696.

[33] R. Mehrotra, N. Xue, and M. Lalmas. 2020. Bandit Based Optimization of Multiple Objectives on a Music Streaming Platform. In *Proc. of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '20)*. ACM, 3224–3233.

[34] A. Moffat and J. Zobel. 2008. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. Inf. Syst.* 27, 1, Article 2 (Dec. 2008), 27 pages.

[35] M. Morik, A. Singh, J. Hong, and T. Joachims. 2020. Controlling Fairness and Bias in Dynamic Learning-to-Rank. In *Proc. of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. ACM, 429–438.

[36] K. P. Murphy. 2021. *Probabilistic Machine Learning: An introduction.* MIT Press.

[37] V. Patil, G. Ghalme, V. Nair, and Y. Narahari. 2020. Achieving Fairness in the Stochastic Multi-Armed Bandit Problem. *Proc. of the AAAI Conference on Artificial Intelligence* 34, 04 (Apr. 2020), 5379–5386.

[38] G. K. Patro, A. Biswas, N. Ganguly, K. P. Gummadi, and A. Chakraborty. 2020. FairRec: Two-Sided Fairness for Personalized Recommendations in Two-Sided Platforms. In *Proc. of The Web Conference (WWW '20)*. ACM, 1194–1204.

[39] J. Pearl. 2009. *Causality.* Cambridge university press.

[40] M. Quadrana, P. Cremonesi, and D. Jannach. 2018. Sequence-Aware Recommender Systems. *ACM Comput. Surv.*, Article Article 66 (July 2018), 36 pages.

[41] S. E. Robertson. 1977. The probability ranking principle in IR. *Journal of documentation* (1977).

[42] M. Rossetti, F. Stella, and M. Zanker. 2016. Contrasting Offline and Online Results when Evaluating Recommendation Algorithms. In *Proc. of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, 31–34.

[43] O. Sakhi, S. Bonner, D. Rohde, and F. Vasile. 2020. BLOB : A Probabilistic Model for Recommendation that Combines Organic and Bandit Signals. In *Proc. of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '20)*. ACM, 783–793.

[44] A. Singh and T. Joachims. 2018. Fairness of Exposure in Rankings. In *Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '18)*. ACM, 2219–2228.

[45] A. Singh and T. Joachims. 2019. Policy Learning for Fairness in Ranking. In *Advances in Neural Information Processing Systems (NeurIPS '19)*, Vol. 32.

[46] N. Sonboli, F. Eskandanian, R. Burke, W. Liu, and B. Mobasher. 2020. Opportunistic Multi-Aspect Fairness through Personalized Re-Ranking. In *Proc. of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20)*. ACM, 239–247.

[47] H. Steck. 2013. Evaluation of Recommendations: Rating-prediction and Ranking. In *Proc. of the 7th ACM Conference on Recommender Systems (RecSys '13)*. ACM, 213–220.

[48] Ö. Sürer, R. Burke, and E. C. Malthouse. 2018. Multistakeholder Recommendation with Provider Constraints. In *Proc. of the 12th ACM Conference on Recommender Systems (RecSys '18)*. ACM, 54–62.

[49] L. Wang, Y. Bai, W. Sun, and T. Joachims. 2021. Fairness of Exposure in Stochastic Bandits. In *International Conference on Machine Learning (ICML'21)*.

[50] H. Yadav, Z. Du, and Thorsten Joachims. 2021. Policy-Gradient Training of Fair and Unbiased Ranking Functions. In *Proc. of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. ACM, 1044–1053.