# Learning to Rank under Multinomial Logit Choice

James A. Grant*[1] and David S. Leslie†[1]

[1]Department of Mathematics and Statistics, Lancaster University, UK

August 17, 2022

**Abstract**

Learning the optimal ordering of content is an important challenge in website design. The learning to rank (LTR) framework models this problem as a sequential problem of selecting lists of content and observing where users decide to click. Most previous work on LTR assumes that the user considers each item in the list in isolation, and makes binary choices to click or not on each. We introduce a multinomial logit (MNL) choice model to the LTR framework, which captures the behaviour of users who consider the ordered list of items as a whole and make a single choice among all the items and a no-click option. Under the MNL model, the user favours items which are either inherently more attractive, or placed in a preferable position within the list. We propose upper confidence bound algorithms to minimise regret in two settings - where the position dependent parameters are known, and unknown. We present theoretical analysis leading to an $\Omega(\sqrt{T})$ lower bound for the problem, an $\tilde{O}(\sqrt{T})$ upper bound on regret for the known parameter version. Our analyses are based on tight new concentration results for Geometric random variables, and novel functional inequalities for maximum likelihood estimators computed on discrete data.

Keywords: Learning to rank; Multinomial Logit choice model; Multi-armed Bandits; Upper Confidence Bound; Concentration Inequalities.

## 1 Introduction

Learning the optimal ordering of content is an important challenge in website design and online advertising. The learning to rank (LTR) framework captures such a challenge via a sequential decision-making model. In this setting, a decision-maker repeatedly selects orderings of items (product advertisements, search results, news articles etc.) and displays them to a user visiting their website. In response the user opts to click on none, one, or more of the displayed items. The objective of the decision-maker will be to maximise the number of clicks received over many iterations of this process. Such an objective is a reasonable and widely-used proxy for the most common interests of a decision-maker in this setting: e.g. maximising profit, and maximising user satisfaction. As such, methods which achieve this objective can be hugely impactful in real-world settings.

*j.grant@lancaster.ac.uk; corresponding author
†d.leslie@lancaster.ac.uk

Recent works (e.g. Kveton et al. (2015), Lagrée et al. (2016)) have considered various formulations of LTR, distinguished by the assumptions on the click model, assumed to govern how users decide to click on items. A majority of previous works utilise a *factored* click model, which assumes, in particular, that the user will click on any displayed item satisfying two conditions: 1) that the user finds the item *attractive*, and 2) the user *examines* the item. The various factored models are differentiated by their specification of the probability of attraction and examination events given particular orderings of content.

Factored models represent the user's choice as a series of binary decisions to click or not click on each examined item. They fail to capture settings where the user's decisions are made among more than two alternatives, for instance choosing between several items considered simultaneously. In this work, we consider LTR under a click model which captures the phenomenon of a user making a *single* decision among *several* alternatives including the option to not click at all. Our click model is based on the *multinomial logit* (MNL) model of user choice (Luce, 1959; Plackett, 1975). We augment the classical MNL model with position effects to capture the relative prominence of different display positions, which may be pre-specified or learned online.

Our model may be more suitable in a variety of settings. For instance, where all items are visible on a screen simultaneously, and are not considered sequentially, or where items are laid out on a grid or more complex display and it is not possible, a priori, to specify a rank order of positions in terms of prominence. This phenomenon has been noticed at a partner company, where the final slot on a homepage has higher click rates than mid-page slots.

### 1.1 Problem Definition

We propose the *Multinomial Logit Learning to Rank* (MNL-LTR) problem. This problem captures the challenge of learning an optimal list of $K$ items among $J$, where the click model is an order-dependent variant of multinomial logit choice.

In each of a series of rounds $t \in [T]$,[1] the decision-maker chooses an *action* $\mathbf{a}_t = (a_{1,t}, \dots, a_{K,t}) \in \mathcal{A} \subset [J]^K$, where $\mathcal{A}$ is the set of all ordered lists of length $K$ consisting of items drawn from $[J]$ without replacement. The action indicates an ordering of $K$ items to display to the user in round $t$. Each action $j \in [J]$ has an associated *attractiveness parameter*, $\alpha_j \in (0, 1]$, and each slot $k \in [K]$ has an associated *position bias* $\lambda_k \in (0, 1]$. We let $\alpha_{k,t} = \alpha_{a_{k,t}}$ refer to the attractiveness parameter of item $a_{k,t}$.

In response to the action $\mathbf{a}_t$, the user will either click on a displayed item or take a no click action. This process is captured via a click variable $Q_t$ taking values in $[K]_0$. The click probabilities follow from the MNL choice model, whose parameters are the products of attractiveness and bias parameters. Specifically, we have

$$P(Q_t = k \mid \mathbf{a}_t) = \frac{\lambda_k \alpha_{k,t}}{1 + \sum_{v=1}^{K} \lambda_v \alpha_{v,t}}, \quad k \in [K], \tag{1}$$

and $P(Q_t = 0 \mid \mathbf{a}_t) = 1 - \sum_{k=1}^{K} P(Q_t = k \mid \mathbf{a}_t)$.

Following the user's choice, the decision-maker receives a reward $R(\mathbf{a}_t) = \mathbb{I}\{Q_t \neq 0\}$. The decision-maker's aim is to maximise their expected cumulative reward over $T$ rounds. The expected reward on an action $\mathbf{a} \in \mathcal{A}$ (in any round) is written, $r(\mathbf{a}) := \mathbb{E}(R(\mathbf{a})) = \sum_{k=1}^{K} P(Q = k \mid \mathbf{a})$. The challenge for the decision maker is that the attractiveness parameters are unknown and the optimal

[1] For an integer $W \geq 1$, we let $[W]$ denote the set $\{1, \dots, W\}$

action is therefore initially unclear. We will consider the problem in two informational settings: where the position biases are known, and unknown.

### 1.2 On Approaches to the Problem

The decision-maker faces a classic exploration-exploitation dilemma and must employ a strategy which balances between reward maximising and information gaining actions. We refer to such a strategy as a *policy*, and formalise it as a (possibly randomised) mapping from a history $\mathcal{H}_{t-1} = \sigma(\mathbf{a}_1, Q_1, \ldots, \mathbf{a}_{t-1}, Q_{t-1})$ to an action $\mathbf{a}_t \in \mathcal{A}$ for each time $t \in [T]$.

We propose *upper confidence bound* (UCB) policies for both the known and unknown position bias settings. UCB approaches are well-studied in the context of multi-armed bandits, following from Lai and Robbins (1985), and are known to achieve optimal regret in a variety of settings. The canonical principle of a UCB approach is as follows. In each round, the policy computes high probability upper confidence bounds on the expected rewards of actions by utilising tight concentration results, and selects an action with maximal associated bound. Intuitively speaking, these approaches are effective because they tend to select actions that either have high reward (and thus are profitable) or high uncertainty (and thus provide a substantial information gain).

In the MNL-LTR setting, the identification of tight concentration results is the most involved aspect of designing UCB policies. In part, this is because the likelihood induced by the MNL model (1) has a complex combinatorial structure, making it hard to identify parameter estimates with known distributional properties. Our proposed strategies subvert this issue by utilising a restriction on the decision-maker's actions such that the likelihood factorises usefully. This technique (first used by Agrawal et al. (2017, 2019)) restricts the decision-maker to repeatedly display any selected ordered list in each round until a no-click event is observed. Unbiased estimators may then be constructed as a sum of geometrically distributed random variables which are functions of the users' stochastic behaviour.

We will be interested in the empirical and theoretical performance of policies, measured in terms of their expected pseudoregret (referred to simply as regret in what follows) in $T$ rounds, defined as,

$$Reg(T) = Tr(\mathbf{a}^*) - \mathbb{E}\bigg(\sum_{t=1}^{T} R(\mathbf{a}_t)\bigg), \tag{2}$$

where $\mathbf{a}^* = \max_{\mathbf{a}\in\mathcal{A}} r(\mathbf{a})$ is an optimal action. Specifically, we will be interested in the order (w.r.t. $T, K$, and $J$) of upper bounds on the regret for our proposed policies, and lower bounds on the regret which hold uniformly across all (reasonable) policies. We will study the problem in two informational settings: one where only attractiveness parameters are unknown, and another where both the attractiveness parameters and position biases are unknown. Our results establish an $\Omega(\sqrt{JKT})$ lower bound on regret, and an upper bound matching this up to logarithmic factors for the known-position bias setting.

### 1.3 Key Contributions

The primary contributions of this work are threefold. Firstly, we provide a new parametric model of LTR based on set-wise user decisions with foundations in classic choice theory.

Second, we derive new theoretical results concerning the concentration of Geometric random variables, giving rise to two new exponential inequalities: The first, an improved high-probability

bound on the sum of non-independent, non-identically distributed (n.i.n.i.d.) Geometric random variables. The second, a high-probability bound on smooth functions of n.i.n.i.d. Geometric random variables, which is applied to the maximum likelihood estimates (MLEs) in the unknown position bias setting to give non-asymptotic confidence sets for all parameters, even in the absence of closed-form expressions for the MLEs.

Finally, based on these results we propose UCB algorithms for the known and unknown position bias settings, and validate their efficacy through derivation of upper and lower bounds on regret - which match up to logarithmic factors - and empirical assessment against other state-of-the-art approaches.

### 1.4 Related Work

A special case of our MNL-LTR model is the MNL bandit (Rusmevichientong et al., 2010). It does not consider ordering of the items and coincides with our model when all position biases are equal.

Initial studies of the MNL-bandit problem presented "explore-then-commit" approaches, which only behave optimally for specific problem classes, and with prior knowledge of certain problem parameters (Rusmevichientong et al., 2010; Sauré and Zeevi, 2013). Agrawal et al. (2019) and Agrawal et al. (2017) since presented UCB and TS approaches to the MNL-bandit respectively, which have $\tilde{O}(\sqrt{JT})$ regret, matching the minimax lower bound derived by Chen and Wang (2018) (up to logarithmic factors). These methods use restrictions on decision-making to permit the construction of estimators with desirable properties. Wang et al. (2018) propose a further approach for the MNL-bandit whose regret is independent of $J$ subject to further assumptions on the attractiveness parameters.

There has since been interest in extending the MNL-bandit model in various directions, considering the best-action identification variant (Chen et al., 2018a), context-dependent variant (Oh and Iyengar, 2019; Chen et al., 2018b), and variants with variable rewards and no 'no-click' event (Bengs and Hüllermeier, 2019; Saha and Gopalan, 2019; Mesaoudi-Paul et al., 2020). None of these works, however, consider the effect of ordering and position biases - i.e. an LTR variant.

Works on LTR are mainly distinguished by different click models (Chuklin et al., 2015), the majority being of the factored form previously described. Two notable choices are the Cascade Model (CM) (Craswell et al., 2008) and the Position Based Model (PBM) (Richardson et al., 2007). Under the CM the user considers each item in sequence, and decides whether or not to click on it before considering any items. If the user clicks an item, or reaches the end of the list without clicking any items, they stop. In contrast, under the PBM, the user may click on multiple items, and chooses whether to examine each item independently, with probabilities similar to our position biases.

Kveton et al. (2015) consider a LTR problem incorporating the CM, and Lagrée et al. (2016) and Komiyama et al. (2017) the PBM. In each of these settings upper confidence bound approaches achieve $O(\sqrt{T})$ regret. Recent works of Zoghi et al. (2017), Lattimore et al. (2018), and Li et al. (2019) have investigated more general click models which include the CM and PBM as special cases. The models of Zoghi et al. (2017) and Li et al. (2019) retain the assumption of a factored model, but are less restrictive than CM, and PBM. The model of Lattimore et al. (2018) makes sufficiently few assumptions to capture a wider range of models, including that which we propose. However such a general approach does not admit as tight theoretical guarantees. Table 1 compares the existing results on regret in LTR and the MNL-bandit with our regret upper bound.

| | MNL choice | LTR | Algorithm | Regret |
|---|---|---|---|---|
| Agrawal et al. (2019) | ✓ | - | UCB | $\tilde{O}\left(\sqrt{JT}\right)$ |
| Agrawal et al. (2017) | ✓ | - | TS | |
| Lattimore et al. (2018) | included as special case | ✓ | TopRank | $O\left(\sqrt{JK^3T}\right)$ |
| **This paper** | ✓ | ✓ | UCB | $\tilde{O}\left(\sqrt{JKT}\right)$ |

Table 1: Comparison of results in the present paper and related work. $T$ denotes the number of rounds, $J$ the number of items, and $K$ the number of items chosen per round.

## 2 Inference

In the MNL-LTR framework, the task of making accurate and efficient inference on the attractiveness parameters is more challenging than in other variants of LTR. Consider the likelihood of the sequence of clicks $Q_{1:T} = \{Q_1, \dots, Q_T\}$ given the attractiveness parameters $\boldsymbol{\alpha}$, position biases $\boldsymbol{\lambda}$, and action sequence $\mathbf{a}_{1:T} = \{\mathbf{a}_1, \dots, \mathbf{a}_T\}$,

$$\mathcal{L}(Q_{1:T} \mid \mathbf{a}_{1:T}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) = \prod_{t=1}^{T} \frac{\sum_{j=0}^{J} \alpha_j \sum_{k=1}^{K} \lambda_k \mathbb{I}\{a_{k,t} = j, Q_t = k\}}{1 + \sum_{j=1}^{J} \alpha_j \sum_{k=1}^{K} \lambda_k \mathbb{I}\{a_{k,t} = j\}}. \tag{3}$$

The likelihood (3) lacks a closed-form maximiser, meaning maximum likelihood estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ can only be computed numerically. Similarly, any Bayesian inference would necessarily be approximate, and computationally intensive. Both of these approximations (which are not necessary in related, factored models) are obstacles to the design and analysis of efficient, optimal sequential decision making policies.

### 2.1 Inference with Known Position Biases

Exact inference is possible if we restrict the manner in which actions are selected. For the MNL-bandit, Agrawal et al. (2019) propose a restriction on decision-making that admits unbiased independent estimators of the attractiveness parameters. Specifically, if each selected set of items is displayed repeatedly until a no-click event is observed, then unbiased estimators of the attractiveness parameters are available. We will show that the same is possible in the MNL-LTR setting, if we display the same ranked list repeatedly until a no-click event occurs.

To describe this approach, we think of the $T$ rounds as being divided into $L \leq T$ epochs of variable length. An epoch $l \in [L]$ will consist of a sequence of consecutive time periods $\mathcal{E}_l \subseteq [T]$. In each epoch $l$ we will offer an ordered list $\mathbf{a}^l \in \mathcal{A}$ repeatedly, until a no-click event is observed. Let $a_k^l$ be the item in position $k$ in epoch $l$ and let $\alpha_k^l$ be the attractiveness parameter of this item, for $k \in [K]$.

As usual in each round $t \in \mathcal{E}_l$ a click variable $Q_t$ is observed. For each slot $k \in [K]_0$ the number of clicks on position $k$ in epoch $l$ is defined as $n_k^l = \sum_{t \in \mathcal{E}_l} \mathbb{I}\{Q_t = k\}$. By the construction of the epochs we always have $n_0^l = 1$, unless $l = L$ and the final epoch is stopped by the completion of the time horizon, rather than a no-click event. We now show that these counts $n_k^l$ can be used to construct simple closed-form estimators of the attractiveness parameters.

The log-likelihood of the observed clicks $\mathbf{n}^l = (n_0^l, n_1^l, \dots, n_K^l)$ in a single epoch $l$, with fixed

action $\mathbf{a}^l$ can be written as,

$$\log \mathcal{L}(\mathbf{n}^l|\mathbf{a}^l, \boldsymbol{\alpha}) = \sum_{k=0}^{K} n_k^l \left[ \log(\lambda_k \alpha_k^l) - \log(1 + \sum_{v=1}^{K} \lambda_v \alpha_v^l) \right].$$

The single-epoch likelihood is maximised by estimators $\hat{\alpha}_k^l = n_k^l/\lambda_k$ for $k \in [K]$.

Inspired by these within-epoch estimators, we may then construct estimators for each attractiveness parameter $\alpha_j$, $j \in [J]$, aggregating over $L$ complete epochs as

$$\bar{\alpha}_j(L) = \frac{\sum_{l=1}^{L} \sum_{k=1}^{K} \mathbb{I}\{a_k^l = j\} n_k^l}{\sum_{l=1}^{L} \sum_{k=1}^{K} \mathbb{I}\{a_k^l = j\} \lambda_k}, \quad j \in [J]. \tag{4}$$

These estimators result from weighted averaging of the within-epoch maximum likelihood estimators - which is preferable to uniform averaging as we should expect epochs where the item $j$ was placed in a slot with a higher position bias to be less variable and thus more reliable. The lemma below, whose proof is reserved for Appendix E, gives the distribution of the random variables $n_k^l$. It follows immediately that our estimators $\hat{\alpha}_k^l$ and $\bar{\alpha}_j(L)$ are unbiased.

**Lemma 1** *For each $k \in [K]$, and $l \in [L]$, $n_k^l$, the number of clicks on the item in position $k$ during epoch $l$, follows an Geometric[2] distribution with parameter $(1 + \lambda_k \alpha_k^l)^{-1}$.*

## 2.2 Inference with Unknown Position Biases

When the position biases are unknown, epoch-based decision making is also useful. In this setting the likelihood is not identified unless we fix one of the position biases, so we fix $\lambda_1 = 1$. This restriction may rescale other parameters, with respect to the known position bias case, but crucially it does not change the interpretation of the model. Some further notation is also useful to describe inference in this setting. Define the $K \times J$ matrix of click counts in $l \in [L]$ epochs as $\mathbf{N}(l)$ having entries,

$$N_{kj} = \sum_{l=1}^{L} \sum_{t \in \mathcal{E}_l} \mathbb{I}\{Q_t = k, \mathbf{a}_k^l = j\}, \quad k \in [K],\ j \in [J].$$

Similarly, define $\tilde{\mathbf{N}}(l)$ as the matrix of counts of selections of item-position combinations, whose entries are

$$\tilde{N}_{kj} = \sum_{l=1}^{L} \mathbb{I}\{\mathbf{a}_k^l = j\}, \quad k \in [K],\ j \in [J].$$

Now, define $\gamma_{jk} = \alpha_j \lambda_k$, $j \in [J]$, $k \in [K]$ to be the products of the attraction probabilities and position biases. Using the known distribution of the click counts $n_k^l$ we can derive an unbiased product parameter estimate $\bar{\gamma}_{jk}(L) = N_{kj}/\tilde{N}_{kj}$ for each $j \in [J]$, and $k \in [K]$. A naive approach would independently estimate the $JK$ product parameters and build UCBs around those. Such an approach does not make efficient use of the data, and as such associated decision-making rules can spend a prohibitively long time exploring, although they will eventually converge to optimal actions. We discuss the limitations of such an approach in more detail in Appendix F and revisit

[2]For clarity, we note that throughout this paper we use the following parametrisation of the geometric distribution. If $X \sim Geom(p)$ then $P(X = x) = (1-p)^x p$, $x \in \mathbb{N} := \{0, 1, \dots\}$.

it in the experiments in Section 6. In the remainder of this section we will focus on direct inference on the attractiveness parameters and position biases.

We may obtain estimates of the attractiveness parameters and position biases via the EM scheme outlined in Algorithm 1. In particular this algorithm exploits that conditioned on estimates of the attractiveness parameters $\hat{\alpha}_{1:J}(L)$ we have an estimate of the position bias for slot $k \in \{2, \ldots, K\}$ as

$$\hat{\lambda}_k(L) = \frac{1}{L} \sum_{j=1}^{J} \frac{\hat{\gamma}_{jk}(L)}{\hat{\alpha}_j(L)} \sum_{l=1}^{L} \mathbb{I}\{\mathbf{a}_k^l = j\} = \frac{1}{L} \sum_{j=1}^{J} \frac{N_{kj}}{\hat{\alpha}_j(L)}. \tag{5}$$

Similarly, we have an estimate of the attractiveness of item $j \in [J]$ given estimates $\hat{\lambda}_{2:K}(L)$ of the position biases, as

$$\hat{\alpha}_j(L) = \frac{1}{\sum_{l=1}^{L} \sum_{k=1}^{K} \mathbb{I}\{\mathbf{a}_k^l = j\}} \sum_{k=1}^{K} \frac{N_{kj}}{\hat{\lambda}_k(L)}, \tag{6}$$

where $\hat{\lambda}_1(L) = \lambda_1 = 1$. Algorithm 1 iterates between estimating position biases and attractiveness parameters until the estimates converge to within some tolerance.[3] The following lemma guarantees the convergence of this EM scheme. Its proof is given in Appendix E, and follows from the unimodality of the log-likelihood function.

**Lemma 2** *The estimators $\boldsymbol{\alpha}^{EM}$, and $\boldsymbol{\lambda}^{EM}$ derived from the EM algorithm, Algorithm 1, converge monotonically to the maximum likelihood estimators.*

---

**Algorithm 1** EM Algorithm for MNL-LTR with Unknown Position Biases

---

**Inputs:** Initial parameter values $\alpha_{j,0}$ for all $j \in [J]$, and $\lambda_{k,0}$ for $k \in \{2, \ldots, K\}$. Tolerance parameter $0 < \xi < 1$. Action and click histories, $\mathbf{a}^{1:L}$, $Q_{1:T}$.

Set $d \leftarrow 1$, $s \leftarrow 0$, and $\lambda_{1,t} \leftarrow 1$ for all $t \geq 0$.

**While** $d > \xi$ **do:**

- Set $s \leftarrow s + 1$.
- **E-Step** For each $k \in \{2, \ldots, K\}$, calculate $\lambda_{k,s}$ according to (5).
- **M-Step** For each $j \in [J]$, calculate $\alpha_{j,s}$ according to (6).
- Calculate $d = \max\left(\max_{k \in \{2,\ldots,K\}} |\lambda_{k,s} - \lambda_{k,s-1}|, \max_{j \in [J]} |\alpha_{j,s} - \alpha_{j,s-1}|\right)$.

**Return** $\lambda_{1,s}, \ldots, \lambda_{K,s}$ and $\alpha_{1,s} \ldots, \alpha_{J,s}$ as estimates of position biases and attractiveness parameters.

---

[3] There is an issue with the numerical stability of this EM scheme, as if a given item or position has no associated clicks, its estimate will go to 0. We can resolve this either by adopting the convention that $0/0 = 0$ or by artificially constraining the estimates to be no smaller than some $\epsilon > 0$

# 3 Concentration Results

In this section we derive concentration results for the parameter estimates in both the known and unknown position bias settings. Quantification of the uncertainty in the parameters is key to designing effective sequential decision-making algorithms, and the results in this section will later be used to construct UCB approaches.

## 3.1 Concentration Results Relevant to the Known Position Bias Setting

As discussed in Section 2.1, the empirical means, $\bar{\alpha}_j$, are weighted averages of Geometric random variables. The following theorem gives a martingale-type concentration result for the sum of geometrically distributed random variables with differing means. This result is not specific to the MNL-LTR or MNL bandit settings, and therefore may be of independent interest.

It is worth noting that the results of Theorem 1 simultaneously have improved coefficients, and a greater generality than alternative results for i.i.d. geometric random variables obtained by Agrawal et al. (2019). We require the greater generality in the MNL-LTR setting because the random variables associated with clicks of an item per epoch will be a) non-identically distributed as they depend on the position bias, and b) non-independent as the assignment of items to slots depends on the previously observed data.

**Theorem 1** *Consider geometric random variables $Y_i$ with parameter $p_i$, $i \in [n]$, where $p_i$ may be a function of $p_1, \ldots, p_{i-1}, Y_1, \ldots Y_{i-1}$. Let $\mu_i = \frac{1-p_i}{p_i}$, and $\sigma_i^2 = \mu_i^2 + \mu_i$. If $\mu_i \leq 1$ for all $i \in [n]$, then we have for all $C > 0$,*

$$P\left(\left|\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \mu_i\right| > \sqrt{2\sum_{i=1}^{n} \sigma_i^2 \log(C) + 4\log(C)}\right) \leq 2C^{-1}, \ \forall n \geq 1. \tag{7}$$

*Furthermore, we have for all $C > 0$,*

$$P\left(\left|\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \mu_i\right| > \sqrt{8\sum_{i=1}^{n} Y_i \log(C) + 4\log(C)} \;\middle|\; A_n\right) \leq 4C^{-1}, \tag{8}$$

*where $A_n = \left\{\sum_{i=1}^{n} \mu_i \geq 8\log(C) + \sqrt{8\sum_{i=1}^{n} \sigma_i^2 \log(C)}\right\}$.*

A full proof of Theorem 1 is provided in Appendix A, but we briefly outline its intuition here. The proof derives a new bound on the central moments of the geometric distribution in order to utilise a Bernstein-like inequality for martingale difference sequences. As the central moments of the geometric distribution lack a closed-form expression, this is non-trivial. We achieve the bound by first bounding the cumulants of the geometric distribution and exploiting a combinatorial link between central moments and cumulants.

The following lemma adapts the result of Theorem 1 to the LTR setting. Its proof is also given in Appendix A. The UCB algorithm we propose in Section 4 for the known position bias setting is designed to exploit these results.

**Lemma 3** *We have for estimators $\bar{\alpha}_j(l)$ $j \in [J]$ defined as in Equation (4), and attractiveness parameters $0 < \alpha_j \leq 1$, $j \in [J]$, the following concentration results, for all $l : \Lambda_{j,l} > 0$*

$$P\left(|\bar{\alpha}_j(l) - \alpha_j| > \sqrt{\frac{4\log(Jl^2/2)}{\Lambda_{j,l}}} + \frac{4\log(Jl^2/2)}{\Lambda_{j,l}}\right) \leq \frac{4}{Jl}, \tag{9}$$

$$P\left(|\bar{\alpha}_j(l) - \alpha_j| > \sqrt{\frac{8\bar{\alpha}_j(l)\log(Jl^2/2)}{\Lambda_{j,l}}} + \frac{8\log(Jl^2/2)}{\Lambda_{j,l}}\right) \leq \frac{6}{Jl}. \tag{10}$$

*Furthermore, for $l : \Lambda_{j,l} > 4\log(Jl^2/2)/\alpha_j$ we have,*

$$P\left(\bar{\alpha}_j(l) > 2\alpha_j + \frac{4\log(Jl^2/2)}{\Lambda_{j,l}}\right) \leq \frac{2}{Jl}. \tag{11}$$

### 3.2 Concentration Results Relevant to the Unknown Position Bias Setting

The derivation of concentration results in the unknown position biases setting is more challenging, since the MLE for any unknown parameter (attractiveness or position bias) does not have a closed form. The asymptotic properties of MLEs are well documented, but there are comparatively few general guarantees relating to finite-time behaviour. However, here we are able to utilise non-asymptotic deviation inequalities on certain functions of random variables to derive concentration properties for a family of MLEs derived from geometrically distributed data.

As in the previous section we have a general concentration result, followed by an application to the MNL-LTR setting. We begin with Theorem 2, whose proof is given in Appendix B which gives a deviation inequality for a function of multivariate Geometric data. The derivation of this result is based on theory from Bobkov and Ledoux (1998) and a logarithmic Sobolev inequality of Joulin and Privault (2004). Before stating our result we introduce a notion of the smoothness of a discrete function expressed in terms of its finite differences.

Let $(\epsilon_{li})_{i\in[d],l\in[n]}$ denote the canonical basis on $\mathbb{R}^{d\times n}$. For a function $f : \mathbb{N}^{d\times n} \to \mathbb{R}$ define the finite difference with respect to the input variable indexed $l, i$,

$$D_{li}f(\mathbf{X}) = F(\mathbf{X} + \epsilon_{li}) - F(\mathbf{X}), \quad \mathbf{X} \in \mathbb{N}^{d\times n}.$$

We say that a function $F : \mathbb{N}^{d\times n} \to \mathbb{R}$ is $(\beta_1, \beta_2)$-smooth, for parameters $\beta_1, \beta_2 > 0$, if,

$$\sum_{l=1}^{n}\sum_{i=1}^{d} |D_{li}F|^2 \leq \beta_1^2, \quad \text{and} \quad \max_{l\in[n]}\max_{i\in[d]}(|D_{li}F|) \leq \beta_2 \quad \forall\, \mathbf{X} \in \mathbb{N}^{d\times n}. \tag{12}$$

**Theorem 2** *Let $n, d \in \mathbb{N}$ and $\mu_l \mid \mu_{1:l-1}$ $l \in [n]$ be a series of conditional multivariate geometric measures on $\mathbb{N}^d$, such that each component, $\mu_{li}$ is a geometric law with parameter $p_{li} \in (0, 1]$ for $i \in [d]$, $l \in [n]$. Define $\mu^n = \bigotimes_{l=1}^n \mu_l$ as the product measure, and let $F$ be a $(\beta_1, \beta_2)$-smooth function with $\beta_1 > 0$, and $\beta_2 \in (0, \max_{i,l}(-\log(1-p_{li}))]$. Then $\mathbb{E}_{\mu^n}(|F|) < \infty$, and for every $\delta > 0$,*

$$\mathbb{P}_{\mu^n}\left(F \geq \mathbb{E}_{\mu^n}(F) + \delta\right) \leq \exp\left(\min\left\{\frac{-\delta^2}{4\beta_1^2 M}, \frac{(\log(1-p))^2\beta_1^2 M}{4\beta_2^2} + \frac{\log(1-p)\delta}{2\beta_2}\right\}\right), \tag{13}$$

*where $M > 0$ is a known finite constant depending on the parameters $\{p_{li}\}_{i\in[d],l\in[n]}$.*

Choosing the function $F$ to be the MLEs in the MNL-LTR setting, we have the following result, giving concentration inequalities for the estimators derived from Algorithm 1.

**Corollary 1** *We have for EM estimators $\alpha_{l,j}^{EM}$, $j \in [J]$ and attractiveness parameters $0 < \alpha_j \leq 1$, $j \in [J]$, that for all $l : \sum_{k=1}^{K} \tilde{N}_{kj} > 0$,*

$$P\left(\left|\alpha_{l,j}^{EM} - \alpha_j\right| > \sqrt{36\beta_{1,l,j}^2 \log(Jl^2)}\right) \leq \frac{2}{Jl^2}.$$

The proof of this corollary is reserved for Appendix B. Its main argument is to recognise that the restriction of $\alpha_l^{EM}$ to its $j^{th}$ output, for fixed selections matrix $\tilde{\mathbf{N}}(l)$, is (subject to a minor rearrangement of the inputs) a function from $\mathbb{N}^{K\times l} \to [0,1]$, to which the functional inequality of Theorem 2 applies.

A similar result will hold for the position bias parameters but is not required. This is since the algorithm we propose in the following section does not need to construct UCBs for the position biases as every slot is utilised in every round.

# 4 Decision-making Algorithms

We now outline our new UCB approaches. As is typical, the algorithms select actions which maximise the expected reward with respect to a set of upper confidence bounds on the attractiveness parameters. Such an optimal action will place the item with the $k^{th}$ largest UCB in the slot with the $k^{th}$ largest position bias (or estimated position bias if position biases are unknown) for each $k \in [K]$. Then, however, the algorithms will repeatedly use this action in each round until a no-click event is observed. This is in contrast to the traditional approach of calculating new UCBs in every round.

**Algorithm 2** Epoch-UCB algorithm for known position biases

Initialise with $l = 0$, and $Q_0 = 0$. Iteratively perform the following for $t \in [T]$,

If $Q_{t-1} = 0$

- Set $l \leftarrow l + 1$
- Calculate UCBs. For $j \in [J]$ compute,

$$\alpha_{j,l}^{UCB} = \bar{\alpha}_j(l-1) + \sqrt{\frac{4\min(1, 2\bar{\alpha}_j(l-1))\log(Jl^2/2)}{\Lambda_{j,l-1}}} + \frac{4\log(Jl^2/2)}{\Lambda_{j,l-1}}.$$

- Select an action $\mathbf{a}_t \in \text{argmax}_{\mathbf{a}\in\mathcal{A}}\, r_{\boldsymbol{\alpha}_l^{UCB}}(\mathbf{a})$ which is optimal with respect to the UCB vector $\boldsymbol{\alpha}_l^{UCB} := (\alpha_{1,l}^{UCB}, \ldots, \alpha_{l,J}^{UCB})$, and observe click variable $Q_t$

otherwise, set action $\mathbf{a}_t = \mathbf{a}_{t-1}$, and observe click variable $Q_t$.

In Algorithm 2 we present our Epoch-UCB method for the variant of MNL-LTR with known position biases. In each epoch $l \in [L]$ the algorithm computes a UCB, $\alpha_{j,l}^{UCB}$ for each item $j \in [J]$.

This UCB is constructed using the concentration results of Lemma 3 to give an upper bound on $\alpha_j$ with high probability. The $\min(1, 2\bar{\alpha}_{j,l-1})$ term allows the UCB to adapt to whichever of (9) and (10) gives the tighter bound.

To state our algorithm for the setting where position biases are unknown define $\alpha^{EM} : \mathbb{N}^{K\times J} \times \mathbb{N}^{K\times J} \to [0,1]^J$ be the function which takes click and selection count matrices as inputs and returns the EM estimates for attractiveness parameters $\boldsymbol{\alpha}$. In effect, $\alpha^{EM}$ represents the application of Algorithm 1.

**Algorithm 3** Epoch-UCB algorithm for unknown position biases

Initialise with $l = 0$ and $Q_0 = 0$. Iteratively perform the following for $t \in [T]$,

If $Q_{t-1} = 0$

- Set $l \leftarrow l + 1$. Compute EM estimators and finite difference bounds,

$$\boldsymbol{\alpha}_l^{EM} = \alpha^{EM}(\mathbf{N}(l-1), \tilde{\mathbf{N}}(l-1)) \tag{14}$$

$$\boldsymbol{\alpha}_{l,kj}^{EM} = \alpha^{EM}(\mathbf{N}(l-1) + \boldsymbol{\epsilon}_{kj}, \tilde{\mathbf{N}}(l-1)), \ \ \forall k, j : \tilde{N}_{kj} \geq 1 \tag{15}$$

$$\beta_{1,l,j}^2 = \sum_{k,s:\tilde{N}_{ks}\geq 1} \left(\alpha_{l-1,j}^{EM} - \alpha_{l-1,ks,j}^{EM}\right)^2 \tilde{N}_{ks}, \ \ \forall j \in [J] \tag{16}$$

- Calculate UCBs. For $j \in [J]$ compute,

$$\alpha_{j,l}^{UCB} = \alpha_{j,l}^{EM} + \sqrt{36\beta_{1,j,L}^2 \log(Jl^2)}.$$

- Select an action $\mathbf{a}_t \in \text{argmax}_{\mathbf{a}\in\mathcal{A}} r_{\boldsymbol{\alpha}_l^{UCB}}(\mathbf{a})$, which is optimal with respect to the UCB vector, and observe click variable $Q_t$.

otherwise, set action $\mathbf{a}_t = \mathbf{a}_{t-1}$, and observe click variable $Q_t$.

In Algorithm 3 we give our policy for the setting where position biases are not known. Its structure is similar to Algorithm 2, but the computation of the UCBs is more involved as it involves finite difference gradients. In each epoch $l \in [L]$, estimates of the MLEs, $\boldsymbol{\alpha}_l^{EM}$, are computed via Algorithm 1 as in (14). Our approach proceeds to calculate further estimates of the attractiveness parameters but on modified data, as in (15). For each item-position pair that has been selected at once, i.e. each $k, j$ with $\tilde{N}_{kj} \geq 1$, we compute $\boldsymbol{\alpha}_{l,kj}^{EM}$, parameter estimates based on $l$ epochs of data but with $N_{kj}$ incremented by 1. A sum of the squared finite differences is then computed as in (16), which is used in the UCB inspired by Corollary 1. This is in place of a supremum bound on the sum of squared differences over all possible outcomes, which would be difficult to compute in practice.

# 5 Regret Bounds

In this section we give upper and lower bounds on the regret for MNL-LTR algorithms. Proposition 1 gives our upper bound on the regret incurred by Algorithm 2 when the position biases are known. Proposition 2 gives a lower bound on the regret of any algorithm, in terms of $S_K = \sum_{k=1}^K \lambda_k$, and

$S_{K,2} = \sum_{k=1}^{K} \lambda_k^2$. The proofs of both results are given in the appendix - Proposition 1 in Appendix C, and Proposition 2 in Appendix D.

**Proposition 1** *The regret in $T$ rounds of the the Epoch-UCB approach, Algorithm 2, for any MNL-LTR problem where the item attractiveness parameters satisfy $\alpha_j \leq \alpha_0 = 1$, $j \in [J]$ and the position biases $\lambda_k \leq 1$, $k \in [K]$ are known satisfies*

$$Reg(T) = O\left(\sqrt{\frac{\log(JT^2)JKT}{\min_{k\in[K]} \lambda_k}}\right).$$

**Proposition 2** *The regret of any algorithm for the MNL-LTR problem with position biases $1 \geq \lambda_1 > \lambda_2 > \cdots > \lambda_K > 0$ satisfying $S_{K,2} > 1$ and $J \geq 4K$ items with attractiveness parameters $\alpha_j \in (0, 1]$, $j \in [J]$, is lower bounded as*

$$Reg(T) = \Omega\left(\sqrt{\frac{JTS_{K,2}^2}{S_K}}\right). \tag{17}$$

The upper and lower bounds in the known position bias case match in their order with respect to $J$ and $T$ (up to logarithmic factors) and match exactly with respect to $K$ when all position biases are 1. Otherwise the gap with respect to $K$ depends on the relative values of the position biases.

For the case of unknown position biases, the more complex form of the UCBs prohibits a full regret analysis. The precise order of the UCB with respect to the the number of times an item has been displayed is unclear, and the usual union bounds over the magnitudes of the UCBs cannot be meaningfully computed. For a lower bound in this setting, we may replace the $S_{K,2}^2/S_K$ term with its largest value and realise an $\Omega(\sqrt{JTK})$ bound.

# 6 Experiments

We now conduct empirical comparisons on three instances of MNL-LTR:

(a) There are $K = 4$ slots with position biases $\boldsymbol{\lambda} = (1, 0.3, 0.2, 0.1)$. There are $J = 6$ items with attractiveness parameters $\boldsymbol{\alpha} = (0.3, 0.28, 0.26, 0.24, 0.22, 0.2)$.

(b) There are $K = 3$ slots with position biases $\boldsymbol{\lambda} = (1, 0.2, 0.9)$. There are $J = 4$ items with attractiveness parameters $\boldsymbol{\alpha} = (0.05, 0.1, 0.15, 0.2)$.

(c) There are $K = 6$ slots with position biases $\boldsymbol{\lambda} = (1, 0.9, 0.7, 0.3, 0.5, 0.7)$ and $J = 30$ items, four having attractiveness parameter 1, two having attractiveness parameter 0.8, and the remaining twenty-four having attractiveness parameter 0.1.

We consider both Algorithm 2 which knows the position biases and Algorithm 3 where the position biases are inferred. We will refer to the former as Epoch-UCB, and the latter as Epoch-UCB UPB (Unknown position biases) in what follows. Experimental results suggest that while the Epoch-UCB UPB algorithm does eventually learn the optimal actions, it can be overly conservative. We therefore also investigate a modification, Epoch-UCB* UPB, which is identical to Algorithm 3 except the UCB for item $j \in [J]$ in epoch $l \in [L]$ is calculated as $\alpha_{j,l}^{UCB} = \alpha_{j,l}^{EM} + 0.5\sqrt{\beta_{1,j,L}^2 \log(\sqrt{J}l)}$.

We compare our algorithms to a range of alternative approaches. Firstly, we have a further known-position-bias epoch-based approach, Epoch-UCB-W. This uses the coefficients we would expect from adapting the weaker concentration inequalities used in Agrawal et al. (2019). Epoch-UCB-W is identical to Algorithm 2 except it calculates UCB index for item $j \in [J]$ in epoch $l \in [L]$ as:

$$\alpha_{j,l}^{UCB} = \bar{\alpha}_{j,l-1} + \sqrt{\frac{48 \min(1, 2\bar{\alpha}_{j,l-1}) \log(\sqrt{J}l/\sqrt{2})}{\Lambda_{j,l-1}}} + \frac{48 \log(\sqrt{J}l/\sqrt{2})}{\Lambda_{j,l-1}}.$$

Second, we also consider the TopRank algorithm of Lattimore et al. (2018). This algorithm can operate without knowledge of the position biases, but assumes that the slots are of decreasing attractiveness. TopRank has a markedly different structure to Epoch-UCB. TopRank maintains a hierarchical partition of the item set, such that the items sit in strata based on their perceived attractiveness. In each round the displayed list is constructed by randomising the order of the $n_1 \geq 1$ items in the top strata and assigning these to the first $n_1$ slots, then randomising the order of the $n_2 \geq 1$ items in the second strata, assigning these to the next $n_2$ slots, and proceeding in such a fashion until all $K$ slots are filled. Items are demoted to lower strata if they have received sufficiently fewer clicks than another item in their strata.

As discussed in Section 1.4, the other LTR approaches that we are aware of are all designed with factored click models in mind, and do not carry performance guarantees to the MNL-LTR setting. We do however investigate the Position Bias Upper Confidence Bound (PBUCB) algorithm of Lagrée et al. (2016), which is based on the position bias click model. This algorithm can use our position bias parameters in terms of its model, but will underestimate the $\alpha$ parameters as it expects that multiple items may be clicked by a user - i.e. its inference model is inconsistent with the MNL data generating process. Further modifications would be necessary to deploy a version of this algorithm if the position biases were not known.

Finally, we compare to the UCB approach described in Appendix F. This approach, which we refer to as 'MNL-bandit' in the figures, ignores some of the LTR structure, and treats the unknown position bias version of the problem as a constrained MNL bandit problem. It learns the product parameters $\gamma_{j,k} = \lambda_k \alpha_j$ individually and avoids the need for running the EM algorithm. As discussed in Appendix F it does have sublinear regret guarantee, but must perform more exploration than other approaches due to being overparameterised.

Note that, problems (b) and (c) give examples where the optimal ordering of items is *not* in decreasing order of attractiveness. In (b) for instance, the final slot, not the second slot, has the second-to-largest position bias. In the unknown position bias variant of this problem, Epoch-UCB UPB and Epoch-UCB* UPB can adapt to this as they actively learn the position biases, but algorithms assuming decreasing position bias, such as TopRank, cannot.

The aforementioned algorithms were applied to problems (a) and (b) over 50000 decision-making rounds, over 40 replications. For problem (c) we use 8000 decision-making rounds, and 40 replications, since the optimal action can be learned more quickly. Figures 1, 2, and 3 display the results, in two forms: the mean regret accumulated through time in their left panes and the distribution of regret in the final rounds in their right. We focus only on the distribution in the final rounds as the results are such that plotting error bars with the each of the seven mean trajectories would make the graphs difficult to read.

Across all three problems we find that our Epoch-UCB and the PBUCB algorithms generally perform best. This is to be expected as they have access to the known position biases and assume a position based model. Our improved coefficients for the known position-bias setting are seen to have

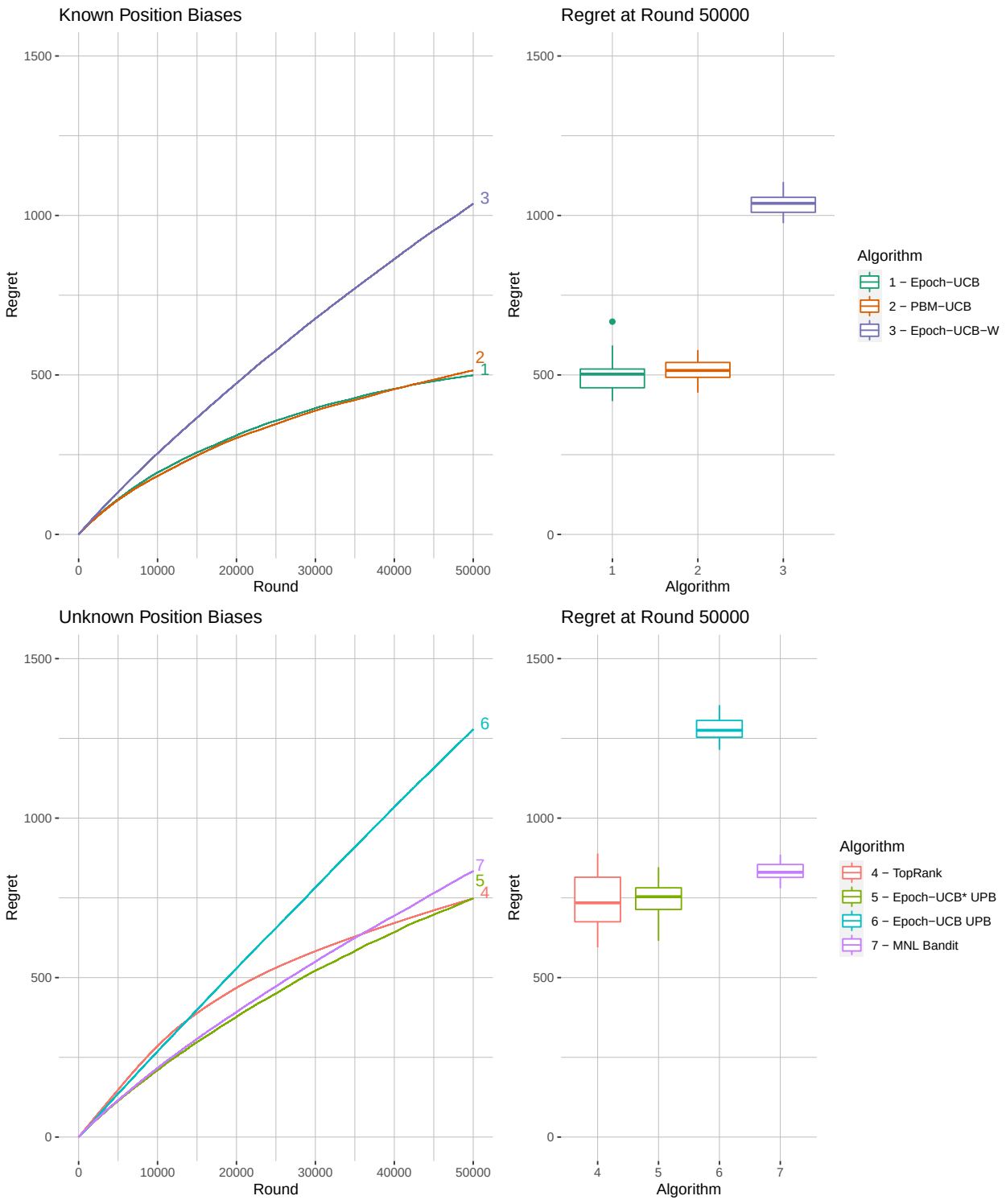


Figure 1: Performance of algorithms on problem (a). The left panel shows the mean cumulative regret trajectory for each algorithm over 50000 rounds. The right panel shows the distribution of regret by algorithm at the end of 50000 rounds.

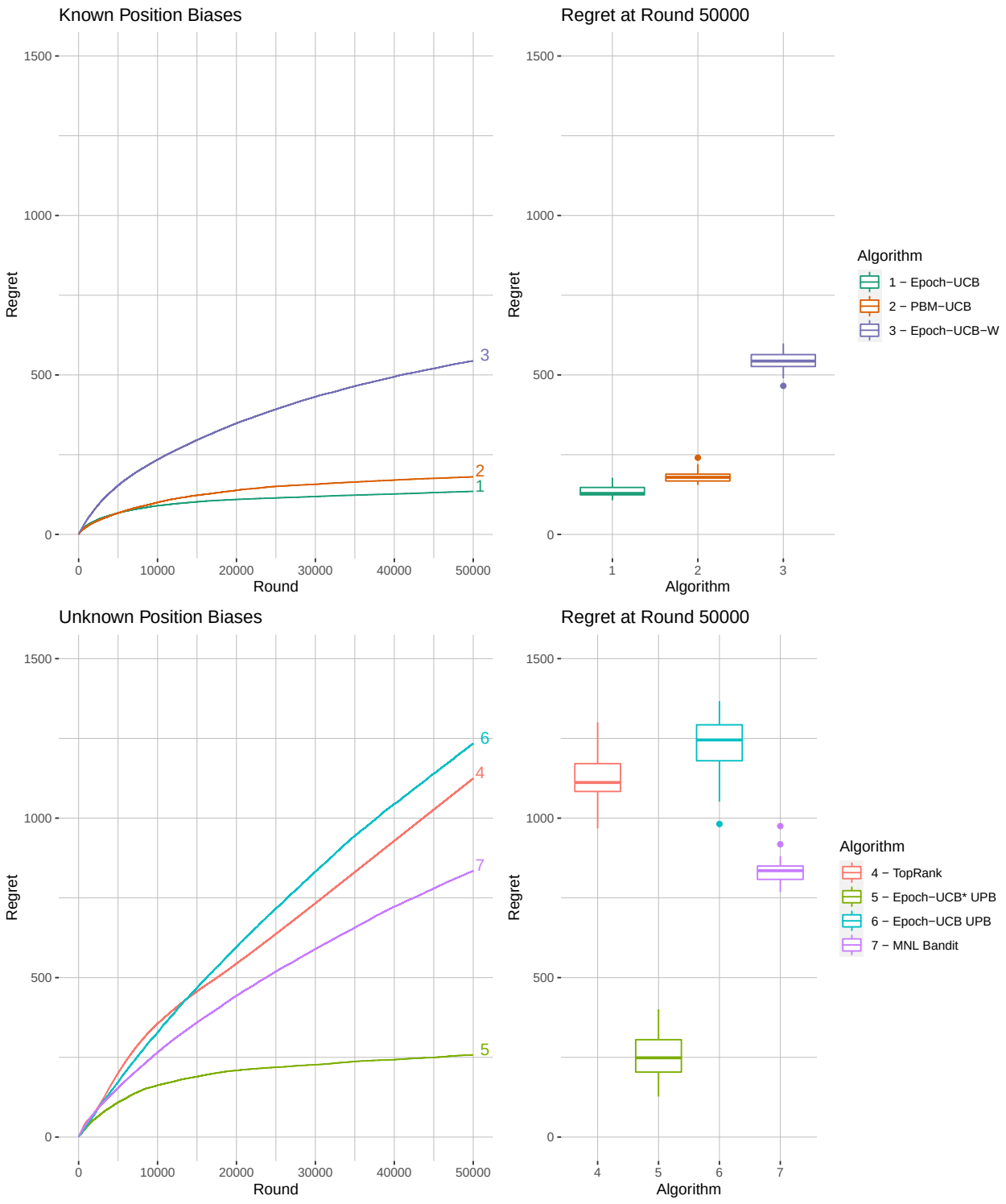


Figure 2: Performance of algorithms on problem (b). The left panel shows the mean cumulative regret trajectory for each algorithm over 50000 rounds. The right panel shows the distribution of regret by algorithm at the end of 50000 rounds.

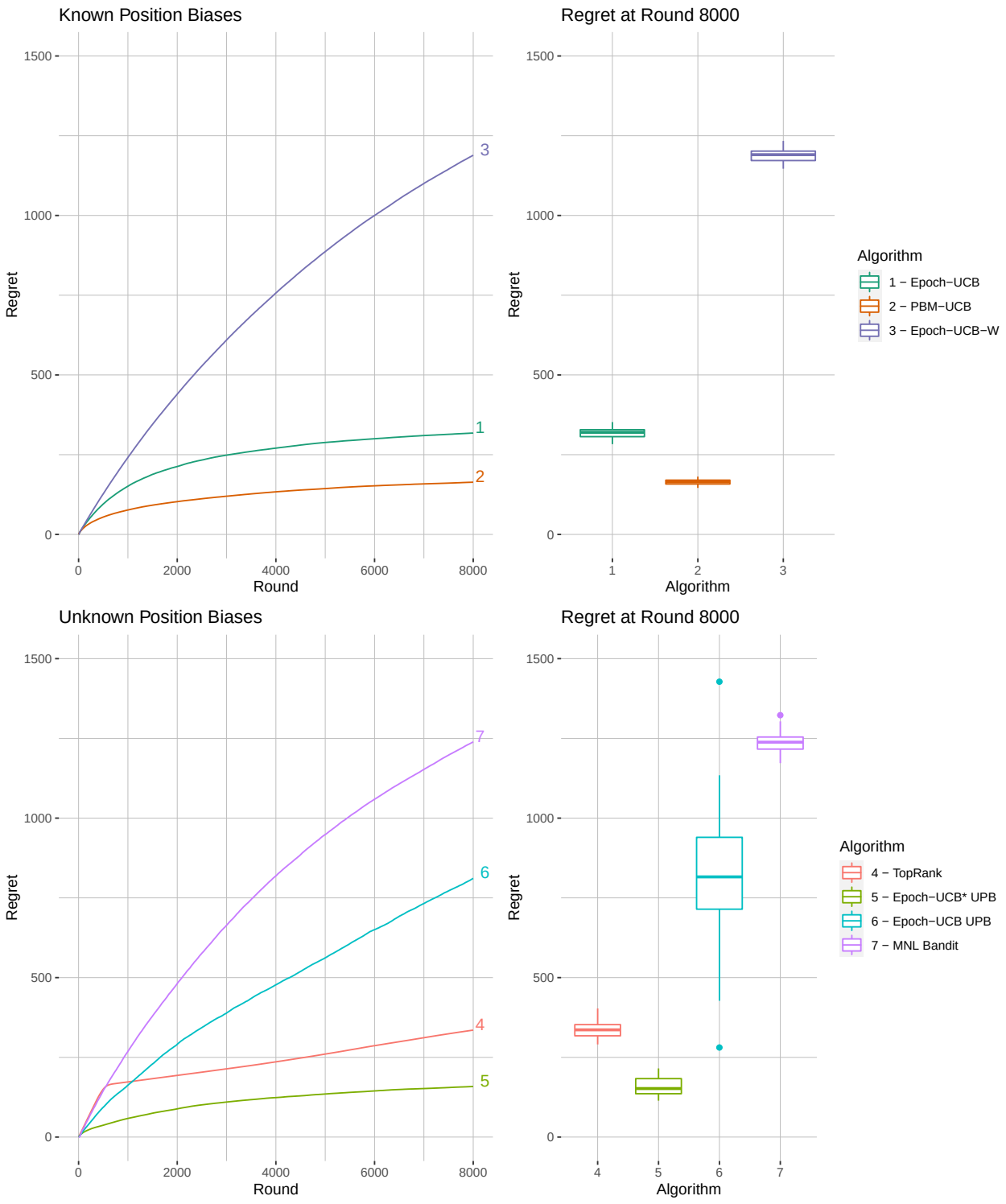


Figure 3: Performance of algorithms on problem (c). The left panel shows the mean cumulative regret trajectory for each algorithm over 50000 rounds. The right panel shows the distribution of regret by algorithm at the end of 8000 rounds.

a substantial benefit as Epoch-UCB-W is much more conservative than Epoch-UCB. Indeed Epoch-UCB-W even suffers worse performance than TopRank which can identify the correct ordering of items despite not knowing the true click model.

In the unknown position bias setting, we see that the unmodified Epoch-UCB UPB approach is also overly conservative, but the approach with modified coefficients Epoch-UCB* UPB is not much worse than the best known position bias algorithms. TopRank does not know the click model but performs well when the position biases are decreasing in the slot index. In problems (b) and (c) where the position biases do not fit this assumption TopRank can identify the top $K$ items reliably, but incurs a linear regret due to repeatedly ordering these suboptimally.

Problem (c) most clearly demonstrates the issue with the MNL-bandit approach. Here, $K$ and $J$ are much larger than in problems (a) and (b), and the MNL-bandit approach continues to explore long after Epoch-UCB* UPB has reached the point of mostly exploiting near-optimal actions. This is due to the fact that the MNL-bandit approach aims to collect data to estimate $JK = 180$ different parameters, whereas Epoch-UCB* UPB utilises the known click model such it only estimates $J + K - 1 = 35$ (assuming $\lambda_1 = 1$). This example displays that although the MNL-bandit approach has a sublinear regret guarantee, as shown in Appendix F, it may be inappropriate in practice.

# 7 Conclusion

In this paper, we have proposed and analysed the multinomial logit choice variant of the learning to rank problem. Distinct from other model-based treatments of learning to rank, our model captures the behaviour of a user who makes a decision over a set of alternatives, rather than making a sequence of independent decisions.

We proposed upper confidence bound approaches for the problem in two informational settings - where the effects of rank are known and unknown respectively. Both of these approaches are derived from new concentration theory. In the known position bias setting, we have derived a verified a Bernstein moment condition on the moments of the Geometric distribution and provided new martingale inequalities for geometric random variables. In the unknown position bias setting we have provided new functional inequalities for geometric data, giving concentration results for numerical maximum likelihood estimators. Further we have provided upper and lower bounds on regret in the known position bias setting, and simulations to display the effectiveness of our approaches.

Our proposed framework makes few assumptions beyond the MNL choice, which is a long-standing popular model in decision theory and may be applicable in numerous domains. Our concentration results are also of potential interest in other areas. Further, we lay a groundwork for further study of MNL-type LTR problems. Future work may consider randomised approaches, utilising similar posterior approximations as in Agrawal et al. (2017), or recently proposed bootstrapping techniques as in Kveton et al. (2019). The development of algorithms for a contextual variant of the problem, or with more complex or alternatively structured actions (perhaps bespoke to particular settings) would also seem to be worthy avenues to follow in the extension of this work.

## References


Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. (2017). Thompson sampling for the mnl-bandit. *arXiv preprint arXiv:1706.00977*.

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. (2019). Mnl-bandit: A dynamic learning approach to assortment selection. *Operations Research*.

Aida, S., Masuda, T., and Shigekawa, I. (1994). Logarithmic sobolev inequalities and exponential integrability. *Journal of Functional Analysis*, 126(1):83–101.

Bengs, V. and Hüllermeier, E. (2019). Preselection bandits under the plackett-luce model. *arXiv preprint arXiv:1907.06123*.

Bobkov, S. G. and Ledoux, M. (1998). On modified logarithmic sobolev inequalities for bernoulli and poisson measures. *Journal of functional analysis*, 156(2):347–365.

Charalambides, C. A. (2002). *Enumerative Combinatorics*. CRC Press.

Chen, X., Li, Y., and Mao, J. (2018a). A nearly instance optimal algorithm for top-k ranking under the multinomial logit model. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2504–2522. SIAM.

Chen, X. and Wang, Y. (2018). A note on a tight lower bound for capacitated mnl-bandit assortment selection models. *Operations Research Letters*, 46(5):534–537.

Chen, X., Wang, Y., and Zhou, Y. (2018b). Dynamic assortment optimization with changing contextual information. *arXiv preprint arXiv:1810.13069*.

Chuklin, A., Markov, I., and Rijke, M. d. (2015). Click models for web search. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 7(3):1–115.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Craswell, N., Zoeter, O., Taylor, M., and Ramsey, B. (2008). An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*, pages 87–94. ACM.

Davies, E. B. and Simon, B. (1984). Ultracontractivity and the heat kernel for schrödinger operators and dirichlet laplacians. *Journal of Functional Analysis*, 59(2):335–395.

de la Peña, V. H. (1999). A general class of exponential inequalities for martingales and ratios. *The Annals of Probability*, 27(1):537–564.

Joulin, A. and Privault, N. (2004). Functional inequalities for discrete gradients and application to the geometric distribution. *ESAIM: Probability and Statistics*, 8:87–101.

Komiyama, J., Honda, J., and Takeda, A. (2017). Position-based multiple-play bandit problem with unknown position bias. In *Advances in Neural Information Processing Systems*, pages 4998–5008.

Kveton, B., Szepesvari, C., Vaswani, S., Wen, Z., Lattimore, T., and Ghavamzadeh, M. (2019). Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 3601–3610.

Kveton, B., Szepesvari, C., Wen, Z., and Ashkan, A. (2015). Cascading bandits: Learning to rank in the cascade model. In *International Conference on Machine Learning*, pages 767–776.

Lagrée, P., Vernade, C., and Cappé, O. (2016). Multiple-play bandits in the position-based model. In *Advances in Neural Information Processing Systems*, pages 1597–1605.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Lattimore, T., Kveton, B., Li, S., and Szepesvari, C. (2018). Toprank: A practical algorithm for online stochastic ranking. In *Advances in Neural Information Processing Systems*, pages 3945–3954.

Li, S., Lattimore, T., and Szepesvari, C. (2019). Online learning to rank with features. In *International Conference on Machine Learning*, pages 3856–3865.

Luce, R. D. (1959). Individual choice behavior.

Mesaoudi-Paul, A. E., Bengs, V., and Hüllermeier, E. (2020). Online preselection with context information under the plackett-luce model. *arXiv preprint arXiv:2002.04275*.

Oh, M.-h. and Iyengar, G. (2019). Thompson sampling for multinomial logit contextual bandits. In *Advances in Neural Information Processing Systems*, pages 3145–3155.

Plackett, R. L. (1975). The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202.

Richardson, M., Dominowska, E., and Ragno, R. (2007). Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*, pages 521–530. ACM.

Rusmevichientong, P., Shen, Z.-J. M., and Shmoys, D. B. (2010). Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations research*, 58(6):1666–1680.

Saha, A. and Gopalan, A. (2019). Combinatorial bandits with relative feedback. In *Advances in Neural Information Processing Systems*, pages 983–993.

Sauré, D. and Zeevi, A. (2013). Optimal dynamic assortment planning with demand learning. *Manufacturing & Service Operations Management*, 15(3):387–404.

Wang, Y., Chen, X., and Zhou, Y. (2018). Near-optimal policies for dynamic multinomial logit assortment selection models. In *Advances in Neural Information Processing Systems*, pages 3101–3110.

Wu, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103.

Zoghi, M., Tunys, T., Ghavamzadeh, M., Kveton, B., Szepesvari, C., and Wen, Z. (2017). Online learning to rank in stochastic click models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4199–4208. JMLR. org.

# A Proof of Geometric Martingale Concentration

In this section, we provide proofs of Theorem 1 and Lemma 3, which comprise the concentration results relevant to the known position bias setting. The following result is a key component of the proof of Theorem 1. It gives a Bernstein-like bound for heavy-tailed martingale data.

**Lemma 4 (Theorem 1.2A of de la Peña (1999))** *Let $\{d_j, \mathcal{F}_j\}$ be a martingale difference sequence with $\mathbb{E}(d_j \mid \mathcal{F}_{j-1}) = 0$, $\mathbb{E}(d_j^2 \mid \mathcal{F}_{j-1}) = \sigma_j^2$, for each $j$, and $V_n^2 = \sum_{j=1}^n \sigma_j^2$. Furthermore assume that*

$$\mathbb{E}(|d_j|^k \mid \mathcal{F}_{j-1}) \leq \frac{k!}{2}\sigma_j^2 c^{k-2} \;\; a.e \tag{18}$$

*or $P(|d_j| \leq c \mid \mathcal{F}_{j-1}) = 1$, for $k > 2$, $0 < c < \infty$. Then for all $x, y > 0$,*

$$P\bigg(\sum_{j=1}^n d_j \geq x, V_n^2 \leq y \text{ for some } n\bigg) \leq \exp\left(\frac{-x^2}{2(y+cx)}\right).$$

## A.1 Proof of Theorem 1

Firstly, we demonstrate that a geometric martingale difference sequence meets the conditions of Lemma 4. Define, $Z_i = \sum_{j=1}^i (Y_i - \mu_i)$ and $W_i = Z_i - Z_{i-1}$. By definition $\{Z_i\}_{i=1}^\infty$ is a martingale and $\{W_i\}_{i=2}^\infty$ is a martingale difference sequence. Immediately, from the distribution of $Y_i$, $i \in [n]$, we have $\mathbb{E}(W_i \mid \mathcal{F}_{i-1}) = 0$ and $\mathbb{E}(W_i^2 \mid \mathcal{F}_{i-1}) = Var(Y_i \mid \mathcal{F}_{i-1}) = \mu_i^2 + \mu_i$.

The higher-order central moments of the Geometric distribution lack a closed-form expression which makes checking condition (18) more complex. Our technique relies on two main steps: we identify a bound on the *cumulants* of the Geometric distribution, and we use a link between the central moments and cumulants from Combinatorics to realise a central moment bound, given in the following lemma.

**Lemma 5** *The central moments $\mu_n$, $n \geq 1$ of the Geometric random variable with parameter $p$ satisfy*

$$\mu_n \leq \frac{!n(1-p)}{p^n}$$

*where $!m$ denotes the number of derangements of an integer $m \geq 1$, defined recursively as*

$$!m = (m-1)(!(m-1)+!(m-2))$$

*where $!0 = 1$ and $!1 = 0$.*

The proof of Lemma 5 is given in Section A.3. It uses the property that the the central moments of any distribution may be expressed in terms of the cumulants $\kappa_k$ of the distribution via *incomplete exponential Bell polynomials*. In particular, we have,

$$\mathbb{E}(W_i^k \mid \mathcal{F}_{i-1}) = \sum_{m=1}^k B_{k,m}(0, \kappa_2, \ldots, \kappa_{k-m+1}), \tag{19}$$

where the summands $B_{k,m}$ are incomplete exponential Bell polynomials (see e.g. Chapter 11 of Charalambides (2002)). For integers $n \geq m \geq 1$ and arguments $x_1, \ldots, x_{n-m+1} \in \mathbb{Z}^{n-m+1}$ these polynomials are defined as

$$B_{n,m}(x_1, \ldots, x_{n-m+1}) = \sum \frac{n!}{j_1! j_2! \ldots j_{n-m+1}!} \left(\frac{x_1}{1!}\right)^{j_1} \left(\frac{x_2}{2!}\right)^{j_2} \cdots \left(\frac{x_{n-m+1}}{(n-m+1)!}\right)^{j_{n-m+1}}, \quad (20)$$

where the sum is over all sequences $j_1, j_2, \ldots, j_{n-m+1}$ of non-negative integers such that $\sum_{i=1}^{n-m+1} j_i = m$ and $\sum_{i=1}^{n-m+1} i j_i = n$.

The bound in Lemma 5 can be adapted to the form required for condition (18). We have the following relationship between the number of derangements and the factorial,

$$!n = \left[\frac{n!}{e}\right] \leq \frac{n!}{2}$$

where $[\cdot]$ is the nearest integer function. It follows that

$$\mathbb{E}(W_i^k | \mathbb{F}_{i-1}) \leq \frac{k!(1-p)}{2p^k} = \frac{k!}{2} \frac{1-p}{p^2} \frac{1}{p^{k-2}}. \quad (21)$$

Thus, the central moments of the Geometric random variable $X_i$ satisfy (18) with $\sigma^2 = \frac{1-p}{p^2}$ and $c = 1/p$.

Thus from Lemma 4 we have, for some $n \geq 1$, and any $x > 0$,

$$P\Big(\sum_{i=1}^n Y_i - \mu_i \geq x\Big) \leq \exp\left(\frac{-x^2}{2\big(\sum_{i=1}^n \sigma_i^2 + x/(\min_i p_i)\big)}\right).$$

Therefore if, for $C > 0$, $x = 2\log(C)/(\min_i p_i) + \sqrt{2\sum_{i=1}^n \sigma_i^2 \log(C)}$, we have

$$\begin{aligned} P\Big(\sum_{i=1}^n Y_i - \mu_i \geq x\Big) &\leq \exp\left(-\frac{\frac{4\log^2(C)}{(\min_i p_i)^2} + \frac{4\sqrt{2\sum_{i=1}^n \sigma_i^2 \log^3(C)}}{\min_i p_i} + 2\sum_{i=1}^n \sigma_i^2 \log(C)}{\frac{4\log(C)}{(\min_i p_i)^2} + \frac{2\sqrt{2\sum_{i=1}^n \sigma_i^2 \log(C)}}{\min_i p_i} + 2\sum_{i=1}^n \sigma_i^2}\right) \\ &\leq \exp\big(-\log(C)\big) = C^{-1} \end{aligned}$$

By symmetry we have the same bound on $P\big(\sum_{i=1}^n \mu_i - Y_i \geq x\big)$, and the statement of (7) follows.

Now consider,

$$P\Big(2\sum_{i=1}^n Y_i \leq \sum_{i=1}^n \mu_i\Big) = P\Big(\sum_{i=1}^n \mu_i - \sum_{i=1}^n Y_i \geq \frac{\sum_{i=1}^n \mu_i}{2}\Big) \leq P\Big(\sum_{i=1}^n \mu_i - \sum_{i=1}^n Y_i \geq \delta \sum_{i=1}^n \mu_i\Big),$$

for any $\delta \in [0, 1/2]$. Choosing

$$\delta = \left(\frac{2\log(C)}{\min_i p_i} + \sqrt{2\sum_{i=1}^n \sigma_i^2 \log(C)}\right) (\sum_{i=1}^n \mu_i)^{-1},$$

and noting $\delta \leq 1/2$ when $\sum_{i=1}^n \mu_i > 4\log(C)/(\min_i p_i) + \sqrt{8\sum_{i=1}^n \sigma_i^2 \log(C)}$, we have by Lemma 4 and a similar manipulation to that used in the proof of (7), that

$$P\left(2\sum_{i=1}^n Y_i \leq \sum_{i=1}^n \mu_i\right) \leq C^{-1}.$$

Furthermore, since $\sigma_i^2 = \mu_i^2 + \mu_i \leq 2\mu_i$ (as $\mu_i \leq 1$), we have also that

$$P\left(4\sum_{i=1}^n Y_i \leq \sum_{i=1}^n \sigma_i^2\right) \leq C^{-1}.$$

It follows that

$$\begin{aligned}
&P\Big(\sum_{i=1}^n Y_i - \mu_i \geq \sqrt{8\sum_{i=1}^n Y_i \log(C)} + \frac{2\log(C)}{\min_i p_i}\Big) \\
&\leq P\Big(\sum_{i=1}^n Y_i - \mu_i \geq \sqrt{2\sum_{i=1}^n \sigma_i^2 \log(C)} + \frac{2\log(C)}{\min_i p_i} \text{ and } 4\sum_{i=1}^n Y_i \leq \sum_{i=1}^n \sigma_i^2\Big) \\
&\leq P\Big(\sum_{i=1}^n Y_i - \mu_i \geq \sqrt{2\sum_{i=1}^n \sigma_i^2 \log(Jl^2/2)} + \frac{2\log(C)}{\min_i p_i}\Big) + P\Big(4\sum_{i=1}^n Y_i \leq \sum_{i=1}^n \sigma_i^2\Big) \leq 2C^{-1}.
\end{aligned}$$

By symmetry we have the high-probability same bound on $\sum_{i=1}^n \mu_i - Y_i$, and the statement of (8) follows. □

## A.2 Proof of Lemma 3

We recall that the number of clicks on item $j \in [J]$ in an epoch $l \in [L]$ is a geometric random variable with parameter $p_{j,l} = \sum_{k=1}^K \mathbb{I}\{a_k^l = j\}(1 + \lambda_k \alpha_j)^{-1}$, as such

$$p_{j,l} \in \left[\left(1 + \max_k \lambda_k \alpha_j\right)^{-1}, \left(1 + \min_k \lambda_k \alpha_j\right)^{-1}\right] \subseteq [0.5, 1]$$

for an epoch where $j \in \mathbf{a}^l$. Thus, the sequence of click counts is a sequence of Geometric random variables of the form considered in Theorem 1. It follows from Theorem 1, specifically equation (7), that for any item $j \in [J]$ the sum of clicks on that item in $L$ epochs obeys,

$$\left|\sum_{l=1}^L \sum_{k=1}^K \mathbb{I}\{a_k^l = j\} n_k^l - \sum_{l=1}^L \sum_{k=1}^K \mathbb{I}\{a_k^l = j\}\lambda_k \alpha_j\right| \leq \sqrt{2\sum_{l=1}^L \sigma_{j,l}^2 \log\left(JL^2/2\right)} + 4\log\left(JL^2/2\right),$$

with probability at least $1 - 4/JL^2$. As per their definition in equation (4) the estimators $\bar{\alpha}_j(l)$ are weighted sums of these click counts, and therefore we have for any $j \in [J]$, $l \in [L]$, and $\mathbf{a}_{1:l}$ such that $\sum_{s=1}^l \sum_{k=1}^K \mathbb{I}\{a_k^s = j\} > 0$,

$$\left|\bar{\alpha}_j(l) - \alpha_j\right| \leq \frac{\sqrt{2\sum_{s=1}^l \sigma_{j,s}^2 \log(Jl^2/2)}}{\sum_{s=1}^l \sum_{k=1}^K \lambda_k \mathbb{I}\{a_k^s = j\}} + \frac{4\log(Jl^2/2)}{\sum_{s=1}^l \sum_{k=1}^K \lambda_k \mathbb{I}\{a_k^s = j\}}, \tag{22}$$

with probability at least $1 - 4/Jl^2$. Notice that

$$\sigma_{j,l}^2 = \mu_{j,l}^2 + \mu_{j,l} \leq 2\mu_{j,l} = 2\alpha_j \sum_{k=1}^{K} \mathbb{I}\{a_k^l = j\} \leq 2 \sum_{k=1}^{K} \mathbb{I}\{a_k^l = j\},$$

and thus we also have, for all $j \in [J]$ and $l \in [L]$,

$$P\left( |\bar{\alpha}_j(l) - \alpha_j| > \sqrt{\frac{4\log(Jl^2/2)}{\sum_{s=1}^{l}\sum_{k=1}^{K} \lambda_k \mathbb{I}\{a_k^s = j\}}} + \frac{4\log(Jl^2/2)}{\sum_{s=1}^{l}\sum_{k=1}^{K} \lambda_k \mathbb{I}\{a_k^s = j\}} \;\middle|\; \mathbf{a}_{1:l} \right) \leq \frac{4}{Jl^2}.$$

Fixing $j$ and $l$ and considering the unconditioned probability, the result stated in equation (9) follows via a union bound.

Similarly, it follows from equation (8) that the data adaptive bound below holds with probability at least $1 - 6/JL^2$,

$$\left| \sum_{l=1}^{L}\sum_{k=1}^{K} \mathbb{I}\{a_k^l = j\} n_k^l - \sum_{l=1}^{L}\sum_{k=1}^{K} \mathbb{I}\{a_k^l = j\} \lambda_k \alpha_j \right| \leq \sqrt{8 \sum_{l=1}^{L}\sum_{k=1}^{K} \mathbb{I}\{a_k^l = j\} n_k^l \log\left(JL^2/2\right) + 4\log\left(JL^2/2\right)},$$

Then by a similar union bound we have the result (10) as stated.

Finally, we consider the probability in equation (11). We have, for $\Lambda_{j,l}$ and $l$ such that $\Lambda_{j,l} > 4\log(Jl^2/2)/\alpha_j$,

$$P\left(\bar{\alpha}_j(l) > 2\alpha_j + \frac{4\log(Jl^2/2)}{\Lambda_{j,l}}\right) \leq P\left(\bar{\alpha}_j(l) - \alpha_j > \sqrt{\frac{4\alpha_j \log(Jl^2/2)}{\Lambda_{j,l}}} + \frac{4\log(Jl^2/2)}{\Lambda_{j,l}}\right) \leq \frac{2}{Jl^2},$$

with the final inequality using Lemma 1 once again. □

### A.3 Proof of Lemma 5

First, we give a recurrence relation for the cumulants of the Geometric distribution. This will be used to verify the Bernstein condition for the central moments. These results may also be of independent interest.

**Lemma 6** *The cumulants $\kappa_n$, $n \geq 1$ of the Geometric random variable with parameter $p$ satisfy,*

$$\kappa_n = \sum_{i=1}^{n} (-1)^{n-i} \frac{h_{n,i}}{p^i} \tag{23}$$

*where the coefficients $h_{n,i}$, $n \geq 1$, $i \leq n$ are recursively defined positive integers satisfying*

$$\begin{aligned} h_{n,1} &= 1 && \forall n \geq 1 \\ h_{n,i} &= i h_{n-1,i} + (i-1) h_{n-1,i-1} && \forall n \geq 3, i \in \{2, \ldots, n-1\} \\ h_{n,n} &= (n-1) h_{n-1,n-1} && \forall n \geq 1. \end{aligned}$$

*Proof:* Firstly we note that the cumulants of the Geometric distribution with parameter $p$ satisfy the recurrence relation

$$\kappa_k = (p-1)\frac{d\kappa_{k-1}}{dp}, \;\; \kappa_1 = \frac{1-p}{p}. \tag{24}$$

The second and third cumulants follow immediately from (24) as

$$\begin{aligned}
\kappa_2 &= (p-1)\frac{d\kappa_1}{dp} = (p-1)\left(\frac{-1}{p^2}\right) = \frac{-1}{p} + \frac{1}{p^2}, \\
\kappa_3 &= (p-1)\frac{d\kappa_2}{dp} = (p-1)\left(\frac{1}{p^2} - \frac{2}{p^3}\right) = \frac{1}{p} - \frac{3}{p^2} + \frac{2}{p^3}.
\end{aligned}$$

Thus we may verify that $\kappa_3$ satisfies the definition in (23). Now assume that $\kappa_n$ matches the definition in (23) for some fixed $n > 3$ and consider $\kappa_{n+1}$. We have from (24) the following,

$$\begin{aligned}
\kappa_{n+1} &= (p-1)\frac{d\kappa_n}{dp} \\
&= (p-1)\sum_{i=1}^{n}(-1)^{n+1-i}\frac{ih_{n,i}}{p^{i+1}} \\
&= \sum_{i=1}^{n}\left[(-1)^{n+1-i}\frac{ih_{n,i}}{p^i} - (-1)^{n+1-i}\frac{ih_{n,i}}{p^{i+1}}\right] \\
&= (-1)^n\frac{h_{n,1}}{p} + \sum_{j=2}^{n}\left[(-1)^{n+1-j}\frac{jh_{n,j}}{p^j} - (-1)^{n-j}\frac{(j-1)h_{n,j-1}}{p^j}\right] - (-1)\frac{nh_{n,n}}{p^{n+1}} \\
&= (-1)^n\frac{h_{n,1}}{p} + \sum_{j=2}^{n}\left[\frac{(-1)^{n+1-j}\big(jh_{n,j} + (j-1)h_{n,j-1}\big)}{p^j}\right] + \frac{nh_{n,n}}{p^{n+1}},
\end{aligned}$$

thus proving the statement by induction. □

Considering this form of the cumulants (23), and the nature of the Bell polynomial (20), it is apparent that the $n^{th}$ central moment $\bar{\mu}_n$ may also be written as $O((1/p)^n)$ polynomials, with some non-negative, integer coefficients $f_{n,1}, \ldots, f_{n,n}$ (to be specified later). Specifically, we may write

$$\bar{\mu}_n = \sum_{i=1}^{n}(-1)^{n-i}\frac{f_{n,i}}{p^i}. \tag{25}$$

Next, we introduce a relevant property of a sequence of non-negative integers, and give a lemma showing that the coefficients of the cumulants and central moments have this property.

**Definition 1 (Alternating Partial Sum (APS))** *A sequence of $n > 0$ non-negative integers, $h_1, \ldots, h_n$ is called APS if for all $k \in [n]$*

$$\sum_{i=1}^{k}(-1)^{n-i}h_i \;\; \begin{cases} \geq 0, \;\; \textit{when } (n-k) \mod 2 = 0, \\ \leq 0, \;\; \textit{when } (n-k) \mod 2 = 1. \end{cases}$$

**Lemma 7** *For any integer $n \geq 3$, and Geometric random variable $X$ with parameter $p$, both the coefficients of the polynomial expression for the cumulants of $X$, $h_{n,1}, \ldots, h_{n,n}$, and the coefficients of the polynomial expression for the central moments of $X$, $f_{n,1}, \ldots, f_{n,n}$ are APS sequences.*

The full proof of Lemma 7 is in the next subsection. The proof has two main steps. First we show that the sequence $h_{n,1}, \ldots, h_{n,n}$ is APS for any $n$ from its recursive formula. Second, we show that the sequence $f_{n,1}, \ldots, f_{n,n}$ can be written as a linear combination of APS sequences (derived from multiplying together cumulants in the Bell polynomial). This linear combination operation preserves the APS property, and thus the sequence $f_{n,1}, \ldots, f_{n,n}$ is also APS.

The proof of Lemma 5 follows from application of Lemma 7. First, we demonstrate that $f_{n,n} =!n$. Recall that the central moments $\mu_n$ are defined in terms of the cumulants as

$$\mu_n = \sum_{m=1}^{n} B_{n,m}(0, \sum_{i=1}^{2}(-1)^{2-i}\frac{h_{2,i}}{p^i}, \ldots, \sum_{i=1}^{n-m+1}(-1)^{n-m+1-i}\frac{h_{n-m+1,i}}{p^i}). \tag{26}$$

It follows that the leading order coefficient $f_{n,n}$ of the polynomial expression for $\mu_n$ can be expressed in terms of incomplete Bell polynomials of the leading order coefficients of the preceding cumulants, i.e.

$$f_{n,n} = \sum_{m=1}^{n} B_{n,m}(0, h_{2,2}, \ldots, h_{n-m+1,n-m+1}) = \sum_{m=1}^{n} B_{n,m}(0, 1!, 2!, \ldots, (n-m)!) \tag{27}$$

where the second equality follows from Lemma 6. The above definition of $f_{n,n}$ coincides with a complete Bell polynomial, such that we have

$$\begin{aligned} f_{n,n} &= B_n(0, 1!, 2!, \ldots, (n-1)!) \\ &= \sum_{\substack{j_2,\ldots,j_n \\ 2j_2+\cdots+nj_n=n}} \frac{n!}{j_2!\ldots j_n!}\left(\frac{1!}{2!}\right)^{j_2}\cdots\left(\frac{(n-1)!}{n!}\right)^{j_n} \\ &= \sum_{\substack{j_2,\ldots,j_n \\ 2j_2+\cdots+nj_n=n}} \frac{n!}{j_2!\ldots j_n!}\left(\frac{1}{2}\right)^{j_2}\cdots\left(\frac{1}{n}\right)^{j_n} =!n. \end{aligned} \tag{28}$$

The final equality follows from the observation that each of the summands in the penultimate expression are the number of permutations in the group of all permutations of $n$ integers with cycle structure $2^{j_2}3^{j_3}\ldots n^{j_n}$. By definition this sum is the number of derangements of $n$.

The second stage of the proof is to demonstrate that $\bar{\mu}_n \leq f_{n,n}/p^n$. First, we note that if $p = 1$ then the Geometric variable $X$ has $P(X = 0) = 1$. Thus, by the definition of $\bar{\mu}_n$ as a central moment, if $p = 1$ then $\bar{\mu}_n = 0$. This implies that the alternating sum of polynomial coefficients $f_{n,1}, \ldots, f_{n,n}$ must be 0 for any $n$, i.e.

$$\sum_{i=1}^{n}(-1)^{n-i}f_{n,i} = 0,$$

and in particular, that

$$f_{n,n} = \sum_{i=1}^{n-1}(-1)^{n-i}f_{n,i}. \tag{29}$$

As $p \leq 1$ by definition, the APS property of $f_{n,1}, \ldots, f_{n-1}$ tells us that

$$\sum_{i=1}^{n-1}(-1)^{n-1-i}\frac{f_{n,i}}{p^i} \geq \frac{1}{p^{n-1}}\sum_{i=1}^{n-1}(-1)^{n-1-i}f_{n,i}. \tag{30}$$

We verify this by considering that

$$\sum_{i=1}^{n-1}(-1)^{n-1-i}f_{n,i} = \sum_{i=1}^{n-1}(-1)^{n-1-i}p^{n-1-i}f_{n,i} + \sum_{i=1}^{n-1-i}(-1)^{n-1-i}(1-p^{n-1-i})f_{n,i}$$
$$\leq \sum_{i=1}^{n-1}(-1)^{n-1-i}p^{n-1-i}f_{n,i},$$

where the inequality holds since the second sum is negative. Dividing both sides by $p^{n-1}$ gives (30).

We then complete the proof by bounding $\bar{\mu}_n$ as follows

$$\bar{\mu}_n = \sum_{i=1}^{n}(-1)^{n-i}\frac{f_{n,i}}{p^i} = \frac{f_{n,n}}{p^n} + \sum_{i=1}^{n-1}(-1)^{n-i}\frac{f_{n,i}}{p^i}$$
$$\leq \frac{f_{n,n}}{p^n} - \frac{f_{n,n}}{p^{n-1}}$$
$$= \frac{f_{n,n}(1-p)}{p^n} = \frac{!n(1-p)}{p^n},$$

where the inequality follows from (30) and (29), and the final equality follows from (28). □

### A.4 Proof of Lemma 7

Firstly we show by induction that the sequences $h_{n,1},\ldots,h_{n,n}$ are APS for all $n \geq 3$.

Consider first the case of $n = 3$. We have, as defined in Lemma 6, that $h_{3,1} = 1$, $h_{3,1} - h_{3,2} = 1-3 = -2$, and $h_{3,1} - h_{3,2} + h_{3,3} = 1-3+2 = 0$. Thus all of the non-negativity and non-positivity conditions are satisfied and the sequence $h_{3,1}, h_{3,2}, h_{3,3}$ is APS. We now assume for some fixed $m \geq 4$ that $h_{m,1},\ldots,h_{m,m}$ is APS, and proceed to consider the sequence $h_{m+1,1},\ldots,h_{m+1,m+1}$.

By definition we have $h_{m,1} = 1$ and $h_{m+1,m+1} = m!$. Thus the APS conditions are satisfied for $k = 1$ and $k = m+1$. We proceed to consider $\sum_{i=1}^{k}(-1)^{m+1-i}h_{m+1,k}$ for $k \in \{2,\ldots,m\}$. We have,

$$\sum_{i=1}^{k}(-1)^{m+1-i}h_{m+1,i} = (-1)^m h_{m,1} + \sum_{i=2}^{k}(-1)^{m+1-i}\big[ih_{m,i} + (i-1)h_{m,i-1}\big]$$
$$= (-1)^{m+1-k}kh_{m,k}. \tag{31}$$

Since all $h_{m,k}$, $m \geq 2, k \leq m$ are positive integers, (31) is positive when $m+1-k \mod 2 = 0$ and negative when $m+1-k \mod 2 = 1$. Thus the APS conditions are satisfied for the sequence $h_{m+1,1},\ldots,h_{m+1,m+1}$ given $h_{m,1},\ldots,h_{m,m}$ is APS. Thus, by induction, the sequences $h_{n,1},\ldots,h_{n,n}$ are APS for all $n \geq 3$.

We next show two properties of APS sequences. Firstly, we have the property that addition preserves APS.

**Property 1 (Preservation of APS under addition)** *If $a_1,\ldots,a_n$ and $b_1,\ldots,b_n$ are APS sequences, then the sequence $a_1+b_1,\ldots,a_n+b_n$ is APS.*

To verify, consider first $j \leq n : n-j \mod 2 = 0$, we have

$$\sum_{i=1}^{j}(-1)^{n-i}(a_i+b_i) = \sum_{i=1}^{j}(-1)^{n-i}a_i + \sum_{i=1}^{j}(-1)^{n-i}b_i \geq 0,$$

since $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ are both APS. Similarly, for $j \leq n : n - j \mod 2 = 1$ we have

$$\sum_{i=1}^{j}(-1)^{n-i}(a_i + b_i) = \sum_{i=1}^{j}(-1)^{n-i}a_i + \sum_{i=1}^{j}(-1)^{n-i}b_i \leq 0,$$

showing that $(a_1 + b_1, \ldots, a_n + b_n)$ are APS.

The second property states that if two polynomials (in the same variable) have APS coefficients, the product of these polynomials has APS coefficients.

**Property 2 (Preservation of APS under polynomial multiplication)** *Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_m$ be APS for $n, m \in \mathbb{N}$, with $\sum_{i=1}^{n}(-1)^{n-i}a_i = 0$ and $\sum_{i=1}^{m}(-1)^{m-i}b_i = 0$. Then, the sequence of polynomial coefficients $c_1, \ldots, c_{n+m}$ such that*

$$\sum_{i=1}^{n+m}(-1)^{n+m-i}c_i x^i = \Big(\sum_{i=1}^{n}(-1)^{n-i}a_i x^i\Big)\Big(\sum_{i=1}^{m}(-1)^{m-i}b_i x^i\Big), \quad x \in \mathbb{R}$$

*are APS.*

To verify this, consider

$$\begin{aligned}
\sum_{i=1}^{n+m}(-1)^{n+m-i}c_i x^i &= \big(a_n x^n - a_{n-1}x^{n-1} + \cdots + (-1)^{n-1}a_1 x\big)\sum_{j=1}^{m}(-1)^{m-j}b_j x_j \\
&= \sum_{j=1}^{m}(-1)^{m-j}b_j\big(a_n x^{n+j} - a_{n-1}x^{n-1+j} + \cdots + (-1)^{n-1}a_1 x^{1+j}\big) \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n+m}(-1)^{n+m-i}d_i^{(j)}x^i
\end{aligned}$$

for coefficients $d_i^{(j)}$, $j \in [m], i \in [n+m]$, defined as follows

$$d_i^{(j)} = \begin{cases} 0, & i < 1 + j \\ a_{i-j}b_j, & i \in \{1 + j, \ldots, n + j\} \\ 0, & i > n + j \end{cases}$$

Now, since $a_1, \ldots, a_n$ is APS and has $\sum_{i=1}^{n}(-1)^{n-i}a_i = 0$, we have

$$\sum_{i=1}^{k}(-1)^{n+m-i}d_i^{(j)} = \begin{cases} 0, & k < 1 + j \\ \sum_{i=1}^{k-j}(-1)^{n-i}a_i, & k \in \{1 + j, \ldots, n + j\} \\ 0, & k > n + j \end{cases}$$

for all $j \in [m]$. Thus the sequence $\{d_i^{(j)}\}_{i=1}^{n+m}$ is APS for each $j \in [m]$ and by Property 1, the coefficients $c_1, \ldots, c_{n+m}$ are also APS.

Using the result that $h_{n,1}, \ldots, h_{n,n}$ is APS for any $n \geq 3$ and the above properties, we will show that the sequences $f_{n,1}, \ldots, f_{n,n}$ are also APS for all $n \geq 3$. Firstly, we recall the definition of the

central moments in terms of cumulants,

$$\bar{\mu}_n = \sum_{k=1}^{n} \sum_{\substack{j_2,\ldots,j_n: \\ j_2+\ldots j_n=k \\ 2j_2+\cdots+nj_n=n}} \frac{n!}{j_2!\ldots j_n!}\left(\frac{\kappa_2}{2!}\right)^{j_2}\left(\frac{\kappa_3}{3!}\right)^{j_3}\cdots\left(\frac{\kappa_n}{n!}\right)^{j_n}.$$

Since $\kappa_2 = O(p^{-2})$, $\kappa_3 = O(p^{-3})$, and $\kappa_n = O(p^{-n})$, each summand is $o(p^{-n})$, and may be written as a polynomial

$$\sum_{i=1}^{n}(-1)^{n-i}\frac{J_i}{p^i} = \frac{n!}{j_2!\ldots j_n!}\left(\frac{\kappa_2}{2!}\right)^{j_2}\left(\frac{\kappa_3}{3!}\right)^{j_3}\cdots\left(\frac{\kappa_n}{n!}\right)^{j_n}$$

where $J_1, \ldots, J_n$ are (possibly zero or negative) coefficients which are a function of variables $j_2, \ldots, j_n$, the order $n$ of the moment in question, and the sequences $h_{m,1}, \ldots, h_{m,n}$ for all $m \leq n$. It follows by Property 2 that the coefficients $J_1, \ldots J_n$ are APS. Finally since the coefficients $f_{n,1}, \ldots, f_{n,n}$ can be written as sums of the $J$ coefficients over the valid combinations of $j_2, \ldots, j_n$, we have by Property 1 that the coefficients $f_{n,1}, \ldots, f_{n,n}$ are APS. □

# B Proof of Maximum Likelihood Concentration

In this section we provide proofs of Theorem 2 and Corollary 1 giving our result on the concentration of the maximum likelihood estimators in the unknown position bias setting.

## B.1 Proof of Theorem 2

The first step of the proof is derive a bound on the relative entropy of functions with respect to the product measure. We make use of the following result from Joulin and Privault (2004). It gives a logarithmic Sobolev inequality tailored to functions of a univariate Geometric distribution.

**Theorem 3 (Theorem 3.7 (Joulin and Privault, 2004))** *Let $\pi$ denote the law of a Geometric random variable with parameter $p$. Let $0 < b < -\log(1-p)$ and let $f : \mathbb{N} \to \mathbb{R}$ such that $|d^+ f| = \max_{k\in\mathbb{N}} |f(k+1) - f(k)| \leq b$ for all $k \in \mathbb{N}$. We have*

$$Ent_\pi\left(e^f\right) \leq \frac{(1-p)e^b}{p(1-\sqrt{(1-p)e^b})}\mathbb{E}_\pi\left(|d^+ f|^2 e^f\right).$$

The chain rule for differential entropy (see e.g. Theorem 8.6.2 of Cover and Thomas (2012)) states that for a series of random variables $X_1, \ldots, X_n$ with joint distribution $\mu^n$,

$$Ent_{\mu^n}(X_1, \ldots, X_n) = \sum_{l=1}^{n} Ent_{\mu_l}(X_l \mid X_1, \ldots, X_{l-1}).$$

Using this result we may extend the univariate bound in Theorem 3 in both dimensions to a bound under product measure. Specially, for every function $G : \mathbb{N}^{d\times n} \to \mathbb{R}$, satisfying $DG_{li} \leq b_{li}$ for constants $0 < b_{li} < -\log(1-p_{li})$ $i \in [d]$, $l \in [n]$, we have

$$Ent_{\mu^n}\left(e^G\right) \leq \max_{i\in[d],l\in[n]} \frac{(1-p_{li})e^{b_{li}}}{p_{li}(1-\sqrt{(1-p_{li})e^{b_{li}}})}\mathbb{E}_{\mu^n}\left(\sum_{i=1}^{d}\sum_{l=1}^{n}|d^+ G_{li}|^2 e^G\right). \tag{32}$$

In particular, under the assumptions of the theorem statement, this holds for $F$, with $\max_{l,i} b_{li} = \beta_2$.

We now follow the so-called Herbst's method (Davies and Simon, 1984; Aida et al., 1994) to achieve a deviation inequality on $F$. For ease in what follows we introduce the function $M_G : \mathbb{R} \to \mathbb{R}$ with

$$M_G(b) = \max_{i\in[d],l\in[n]} \frac{(1-p_{li})e^b}{p_{li}(1-\sqrt{(1-p_{li})e^b})}, \quad b \in \left(0, \max_{i\in[d],l\in[n]} -\log(1-p_{li})\right). \tag{33}$$

Further we let $\bar{p}$ be the parameter attaining the maximum in (33), i.e.

$$\bar{p} = \operatorname*{argmax}_{p_{li}:i\in d,l\in[n]} \frac{(1-p_{li})e^b}{p_{li}(1-\sqrt{(1-p_{li})e^b})},$$

for any valid $b$.

Applying (32) to $\eta F$ for every $0 < \eta \leq -\log(1-\bar{p})/(2\beta_2)$ and substituting the definition for entropy we have,

$$\mathbb{E}_{\mu^n}\left(\eta F e^{\eta F}\right) - \mathbb{E}_{\mu^n}\left(e^{\eta F}\right)\log\left(\mathbb{E}_{\mu^n}\left(e^{\eta F}\right)\right) \leq M_G(\eta\beta_2)\mathbb{E}_{\mu^n}\left(\sum_{l=1}^{n}\sum_{i=1}^{d_1}\eta^2|F(\mathbf{X}+\epsilon_{li}) - F(\mathbf{X})|^2 e^{\eta F}\right).$$

Exploiting the assumed bound (12), and introducing the notation $H(\eta) = \mathbb{E}_{\mu^n}\left(e^{\eta F}\right)$ we may then rewrite the above display as,

$$\eta H'(\eta) - H(\eta)\log\left(H(\eta)\right) \leq \eta^2\beta_1^2 M_G(\eta\beta_2)H(\eta).$$

We then set $K(\eta) = \frac{1}{\eta}\log(H(\eta))$, with $K(0) = H'(0)/H(0) = \mathbb{E}_{\mu^n}(F)$ and observe,

$$K'(\eta) \leq \frac{\eta^2\beta_1^2 M_G(\eta\beta_2)H(\eta)}{\eta^2 H(\eta)} = \beta_1^2 M_G(\eta\beta_2) \leq \beta_1^2 M_G\left(\frac{-\log(1-\bar{p})}{2}\right)$$

since $M_G$ is an increasing function. We will define $M := M_G\left(\frac{-log(1-\bar{p})}{2}\right)$ for convenience. As such we may bound $K(\eta) \leq \mathbb{E}_{\mu^n}(F) + \eta\beta_1^2 M$, and derive the exponential inequality,

$$\mathbb{E}_{\mu^n}\left(e^{\eta F}\right) \leq \exp\left(\eta\mathbb{E}_{\mu^n}\left(F\right) + \eta^2\beta_1^2 M\right), \quad 0 < \eta \leq \frac{-\log(1-\bar{p})}{2\beta_2}. \tag{34}$$

Finally, we apply a Chernoff bound to $F$, and substitute (34), giving,

$$\mathbb{P}_{\mu^n}\left(F \geq \mathbb{E}_{\mu^n}(F) + \delta\right) \leq \exp\left(\eta\mathbb{E}_{\mu^n}\left(F\right) + \eta^2\beta_1^2 M - \eta\mathbb{E}_{\mu^n}(F) - \eta\delta\right)$$

which when minimised over $\eta \in (0, -\log(1-\bar{p})/(2\beta_2)]$, yields the stated result. □

## B.2 Proof of Corollary 1

The function $\alpha^{EM}$ which computes the EM estimates of $\boldsymbol{\alpha}$, is posed as a function from $\mathbb{N}^{K\times J}\times\mathbb{N}^{K\times J}$ to $[0,1]^J$, where the input matrices are of the form of $\mathbf{N}(L)$ and $\tilde{\mathbf{N}}(L)$. Recall that we have $\alpha^{EM}(\mathbf{N}(L), \tilde{\mathbf{N}}(L)) = \boldsymbol{\alpha}^{EM}$ where $\boldsymbol{\alpha}^{EM}$ are the EM estimates of $\boldsymbol{\alpha}$ derived from Algorithm 1. For the purposes of this proof we will define $\bar{\alpha}^{EM} : \mathbb{N}^{K\times L}\times[J]^{K\times L} \to [0,1]^J$, which computes the same $\boldsymbol{\alpha}^{EM}$ estimates but via an alternative arrangement of the input data.

Let $\bar{\mathbf{N}}(L)$ be the $K \times L$ matrix whose $l^{th}$ column ($l \in [L]$) is $\mathbf{n}^l$, which we recall is the vector of clicks per slot in epoch $l$. Let $\bar{\mathbf{A}}(L)$ be the $K \times L$ matrix whose $l^{th}$ column ($l \in [L]$) is $\mathbf{a}^l$ the action vector in epoch $l$. Define $\bar{\alpha}^{EM}$ such that $\bar{\alpha}^{EM}(\bar{\mathbf{N}}(L), \bar{\mathbf{A}}(L)) = \alpha^{EM}(\mathbf{N}(L), \tilde{\mathbf{N}}(L)) = \boldsymbol{\alpha}^{EM}$.

Then, for fixed $\bar{\mathbf{A}}(L)$, the restriction of $\bar{\alpha}^{EM}$ to its $j^{th}$ output, $\bar{\alpha}_j^{EM}$, is, a function from $\mathbb{N}^{K\times L}$ to $[0,1]$. It also follows from the definitions above that $\mathbb{E}(\bar{\alpha}_j^{EM}) = \alpha_j$. Finally since the entries of the (non-fixed) input matrix $\bar{\mathbf{N}}(L)$ are Geometric random variables, the function $\bar{\alpha}_j^{EM}$ fits within the framework of Theorem 2. We therefore have that

$$P\left(|\alpha_j^{EM}(\mathbf{N}(L), \tilde{\mathbf{N}}(L)) - \alpha_j| \geq \delta\right) = P\left(|\bar{\alpha}_j^{EM}(\bar{\mathbf{N}}(L), \bar{\mathbf{A}}(L)) - \alpha_j| \geq \delta\right) \leq \exp\left\{\frac{-\delta^2}{4\beta_{1,L,j}^2 M}\right\}$$

where

$$\beta_{1,j,l} = \sup_{\mathbf{X}\in\mathbb{N}^{K\times L}} \sqrt{\sum_{l=1}^{L}\sum_{k=1}^{K}\left|\bar{\alpha}_j^{EM}(\mathbf{X} + \epsilon_{lk}) - \bar{\alpha}_j^{EM}(\mathbf{X})\right|^2} \tag{35}$$

for any $\mathbf{X} \in \mathbb{N}^{K\times L}$.

We recall that $M = M_G(\max_{k\in[K],l\in[L]} - \log(1-p_{kl})/2)$. Specifically in the context of our MNL-LTR problem, we have $p_{kl} \in [1/2, 1)$ for all $k \in [K]$, $l \in [L]$. Thus, $M \leq \sqrt{0.5}/(0.5(1-(0.5)^{1/4})) \leq 9$. Rearranging the exponential inequality and substituting a bound on $M$ we therefore have,

$$P\left(|\alpha_j^{EM}(\mathbf{N}(L), \tilde{\mathbf{N}}(L)) - \alpha_j| \geq \sqrt{72\beta_{1,L,j}^2 \log(\sqrt{J}L)}\right) \leq \frac{2}{JL^2}. \quad \square$$

### B.3 Bound on Finite Difference Parameter

In this section, we derive a bound on the sensitivity parameter $\beta_{1,j,l}$ which enables an analysis of the regret. We first derive an alternative expression for $\beta_{1,j,l}$ to that in (35) in terms of the $\alpha^{EM}$ functions, as oppose to the $\bar{\alpha}^{EM}$ functions, as

$$\beta_{1,j,l} = \sup_{\mathbf{N}\in\mathbb{N}^{K\times J}} \sqrt{\sum_{i=1}^{J}\sum_{k=1}^{K}\tilde{N}_{ki}\left|\alpha_j^{EM}(\mathbf{N}, \tilde{\mathbf{N}}) - \alpha_j^{EM}(\mathbf{N} + \epsilon_{ki}, \tilde{\mathbf{N}})\right|^2}.$$

Bounding the finite differences with partial derivatives, we have the following bound on the sensitivity parameter,

$$\beta_{1,j,l} \leq \sup_{\mathbf{N}\in\mathbb{N}^{K\times J}} \sqrt{\sum_{i=1}^{J}\sum_{k=1}^{K}\tilde{N}_{ki} \max_{n_{ki}\in[N_{ki},N_{ki}+1]}\left(\frac{\partial \alpha_j^{EM}}{\partial N_{ki}}\bigg|_{N_{ki}=n_{ki}}\right)^2}. \tag{36}$$

In what follows we shall obtain a further analytical bound on this expression in order to relate the sensitivity parameter to the selected actions and number of rounds.

For the purposes of this section, let $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\lambda}_{-1})$ denote the length $J + K - 1$ vector of the unknown parameters in the unknown position biases model. As before, $\mathbf{N}$ and $\tilde{\mathbf{N}}$ are $J \times K$ matrices of click- and play-counts for each item-slot combination. We will suppress dependence on $L$ for brevity in this section. In order to express gradients with respect to click counts conveniently, we introduce the notation $\mathbf{n}$ to represent a vectorised version of $\mathbf{N}$. Here element $m$ of $\mathbf{n}$ corresponds to element $N_{(m \mod K),\lceil m/K\rceil}$ of the matrix $\mathbf{N}$.

For a given pair $\mathbf{n}, \tilde{\mathbf{N}}$, the maximum likelihood estimate,

$$\boldsymbol{\theta}^*(\mathbf{n}, \tilde{\mathbf{N}}) \in \underset{\boldsymbol{\theta} \in [0,1]^{J+K-1}}{\operatorname{argmax}} \ell\left(\boldsymbol{\theta}; \mathbf{n}, \tilde{\mathbf{N}}\right),$$

is found where $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{n}, \tilde{\mathbf{N}}) = 0$. It then follows from the continuous differentiability of the likelihood, and an application of the Implicit Function theorem that the gradient of the maximiser with respect to the click vector $\mathbf{n}$, within an open neighbourhood of a particular solution $\boldsymbol{\theta}_0^*$ may be written,

$$\nabla_{\mathbf{n}} \boldsymbol{\theta}^*(\mathbf{n}, \tilde{\mathbf{N}}) = -\left[\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell\left(\boldsymbol{\theta}; \mathbf{n}, \tilde{\mathbf{N}}\right)\right]^{-1} \nabla_{\boldsymbol{\theta}\mathbf{n}} \ell\left(\boldsymbol{\theta}; \mathbf{n}, \tilde{\mathbf{N}}\right). \tag{37}$$

Elements of $\nabla_{\mathbf{n}} \boldsymbol{\theta}^*$, a $(J+K-1) \times JK$ matrix provide the gradient of maximum likelihood estimators with respect to particular click counts. We will use these to derive a bound on $\beta_{1,j,L}$.

The gradient vector $\nabla_{\boldsymbol{\theta}} \ell$ has entries,

$$\frac{\partial \ell}{\partial \alpha_j} = \sum_{k=1}^{K} \frac{N_{kj}}{\alpha_j} - \frac{\lambda_k (N_{kj} + \tilde{N}_{kj})}{(1 + \alpha_j \lambda_k)}, \quad \frac{\partial \ell}{\partial \lambda_k} = \sum_{j=1}^{J} \frac{N_{kj}}{\lambda_k} - \frac{\alpha_j (N_{kj} + \tilde{N}_{kj})}{(1 + \alpha_j \lambda_k)}.$$

and the Hessian $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell$ contains second derivatives of the following form

$$\frac{\partial^2 \ell}{\partial \alpha_j^2} = \sum_{k=1}^{K} \frac{\lambda_k^2 (N_{kj} + \tilde{N}_{kj})}{(1 + \alpha_j \lambda_k)^2} - \frac{N_{kj}}{\alpha_j^2}, \quad \frac{\partial^2 \ell}{\partial \lambda_k^2} = \sum_{j=1}^{J} \frac{\alpha_j^2 (N_{kj} + \tilde{N}_{kj})}{(1 + \alpha_j \lambda_k)^2} - \frac{N_{k,j}}{\lambda_k^2}, \quad \frac{\partial^2 \ell}{\partial \alpha_j \lambda_k} = -\frac{(N_{kj} + \tilde{N}_{kj})}{(1 + \alpha_j \lambda_{kj})^2},$$

with $\partial^2 \ell / \partial \alpha_j \partial \alpha_j' = 0$ where $j \neq j'$ and $\partial^2 \ell / \partial \lambda_k \partial \lambda_k' = 0$ where $k \neq k'$. Finally, the elements of $\nabla_{\boldsymbol{\theta}\mathbf{n}}$ are the second derivatives,

$$\frac{\partial^2 \ell}{\partial \alpha_j \partial N_{kj}} = \frac{1}{\alpha_j} - \frac{\lambda_k}{(1 + \alpha_j \lambda_k)}, \quad \frac{\partial^2 \ell}{\partial \lambda_k \partial N_{kj}} = \frac{1}{\lambda_k} - \frac{\alpha_j}{(1 + \alpha_j \lambda_k)}, \tag{38}$$

and $\partial^2 \ell / \partial \alpha_j N_{kj'} = 0$ for $j \neq j'$ and $\partial^2 \ell / \lambda_k N_{k'j} = 0$ for $k \neq k'$.

It follows that the gradient of a particular attractiveness parameter estimate $\alpha_j^*$ with respect to a particular click count $N_{kl}$, with $k \geq 2$, may be written,

$$\begin{aligned} \frac{\partial \alpha_j^*}{\partial N_{kl}} &= -\sum_{s=1}^{J} \left(\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\right)_{js} \frac{\partial^2 \ell}{\partial \alpha_s \partial N_{kl}} - \sum_{m=2}^{K} \left(\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\right)_{j,J+m-1} \frac{\partial^2 \ell}{\partial \lambda_m \partial N_{kl}} \\ &= \begin{cases} -\left(\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\right)_{jj} \frac{\partial^2 \ell}{\partial \alpha_j \partial N_{kj}} - \left(\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\right)_{j,J+k-1} \frac{\partial^2 \ell}{\partial \lambda_k \partial N_{kj}}, & \text{if } j = l \\ -\left(\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\right)_{j,J+k-1} \frac{\partial^2 \ell}{\partial \lambda_k \partial N_{kl}}, & \text{if } j \neq l. \end{cases} \end{aligned} \tag{39}$$

where $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}$ denotes the inverse Hessian. In the case of leading slot, $k = 1$, with fixed attractiveness parameter $\lambda_1 = 1$, we have gradient,

$$\frac{\partial \alpha_j^*}{\partial N_{1l}} = \begin{cases} -\left(\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\right)_{jj} \frac{\partial^2 \ell}{\partial \alpha_j \partial N_{1j}}, & \text{if } j = l \\ 0, & \text{if } j \neq l. \end{cases} \tag{40}$$

We proceed to bound these gradients uniformly, and apply said bounds to derive a bound on $\beta_{1,j,l}$.

By definition of the matrix inverse, utilising the sparsity of $\nabla_{\boldsymbol{\theta\theta}}$, and with a view to evaluating (37) we may write elements of the row vector $(\nabla^{-1}_{\boldsymbol{\theta\theta}})_{j\cdot}$, for any $j \in [J]$, as solutions to the system of equations,

$$\begin{aligned}
\left(\nabla^{-1}_{\boldsymbol{\theta\theta}}\right)_{jj} \frac{\partial^2 \ell}{\partial \alpha_j^2} + \sum_{k=2}^{K} \left(\nabla^{-1}_{\boldsymbol{\theta\theta}}\right)_{j,J+k-1} \frac{\partial^2 \ell}{\partial \lambda_k \partial \alpha_j} &= 1, \\
\left(\nabla^{-1}_{\boldsymbol{\theta\theta}}\right)_{ji} \frac{\partial^2 \ell}{\partial \alpha_i^2} + \sum_{k=2}^{K} \left(\nabla^{-1}_{\boldsymbol{\theta\theta}}\right)_{j,J+k-1} \frac{\partial^2 \ell}{\partial \lambda_k \partial \alpha_i} &= 0, \quad i \in [J] \setminus \{j\}, \\
\left(\nabla^{-1}_{\boldsymbol{\theta\theta}}\right)_{j,J+k-1} \frac{\partial^2 \ell}{\partial \lambda_k^2} + \sum_{s=1}^{J} \left(\nabla^{-1}_{\boldsymbol{\theta\theta}}\right)_{j,s} \frac{\partial^2 \ell}{\partial \lambda_k \partial \alpha_s} &= 0, \quad k \in [K] \setminus \{1\}.
\end{aligned}$$

So long as this system is non-singular, the linearity of the derivatives in play counts $\tilde{N}_{\cdot\cdot}$ can be used to establish the order of the inverse Hessian terms. In particular, we have,

$$\left|\frac{\partial^2 \ell}{\partial \alpha_j^2}\right| = \Omega\left(\tilde{N}_{1j}, \tilde{N}_{2j}, \dots \tilde{N}_{Kj}\right), \ j \in [J], \qquad \left|\frac{\partial^2 \ell}{\partial \lambda_k^2}\right| = \Omega\left(\tilde{N}_{k1}, \tilde{N}_{k2}, \dots \tilde{N}_{kJ}\right), \ k \in [K] \setminus \{1\},$$

$$\left|\frac{\partial^2 \ell}{\partial \lambda_k \partial \alpha_j}\right| = \Omega\left(\tilde{N}_{kj}\right), \ j \in [J], \ k \in [K] \setminus \{1\}.$$

It follows, since there are more elements of $\tilde{\mathbf{N}}$ than of the row vector $(\nabla^{-1}_{\boldsymbol{\theta},\boldsymbol{\theta}})_{j\cdot}$ that,

$$\begin{aligned}
\left|\left(\nabla^{-1}_{\boldsymbol{\theta\theta}}\right)_{ji}\right| &= O\left(\tilde{N}_{1i}^{-1}, \tilde{N}_{2i}^{-1}, \dots, \tilde{N}_{Ki}^{-1}\right), \ j \in [J], \quad i \in [J], \\
\left|\left(\nabla^{-1}_{\boldsymbol{\theta\theta}}\right)_{j,J+k-1}\right| &= O\left(\tilde{N}_{k1}^{-1}, \tilde{N}_{k2}^{-1}, \dots, \tilde{N}_{kJ}^{-1}\right), \ j \in [J], \quad k \in [K] \setminus \{1\}.
\end{aligned}$$

Expanding the terms of the bound (36), using the expressions (39) and (40), we have

$$\begin{aligned}
\beta_{i,j,L} \leq \sup_{\mathbf{N} \in \mathbb{N}^{K \times J}} \Bigg( & \tilde{N}_{1j} \max_{n_{1j} \in [N_{1j}, N_{1j}+1]} \left[ - \left(\nabla^{-1}_{\boldsymbol{\theta\theta}}\right)_{jj} \frac{\partial^2 \ell}{\partial \alpha_j \partial N_{1j}} \right]^2 \\
& + \sum_{i \neq j} \sum_{k=2}^{K} \tilde{N}_{ki} \max_{n_{ki} \in [N_{ki}, N_{ki}+1]} \left[ - \left(\nabla^{-1}_{\boldsymbol{\theta\theta}}\right)_{j,J+k-1} \frac{\partial^2 \ell}{\partial \lambda_k \partial N_{ki}} \right]^2 \\
& + \sum_{k=2}^{K} \tilde{N}_{kj} \max_{n_{kj} \in [N_{kj}, N_{kj}+1]} \left[ - \left(\nabla^{-1}_{\boldsymbol{\theta\theta}}\right)_{jj} \frac{\partial^2 \ell}{\partial \alpha_j \partial N_{kj}} - \left(\nabla^{-1}_{\boldsymbol{\theta\theta}}\right)_{j,J+k-1} \frac{\partial^2 \ell}{\partial \lambda_k \partial N_{kj}} \right]^2 \Bigg)^{1/2}.
\end{aligned}$$

Since the second derivatives which mix parameter and click derivatives (those of the form $\partial^2 \ell / \partial \alpha_{\cdot} . \partial N_{\cdot\cdot}$ and $\partial^2 \ell / \partial \lambda_{\cdot} . \partial N_{\cdot\cdot}$ defined in (38)) are independent of $\mathbf{N}$ and $\tilde{\mathbf{N}}$, there exists a constant $C > 0$ such that,

$$\beta_{1,j,L} \leq C \left( \sum_{k=1}^{K} \tilde{N}_{kj} \frac{1}{(\max_k \tilde{N}_{kj})^2} + \sum_{i=1}^{J} \sum_{k=1}^{K} \tilde{N}_{ki} \frac{1}{(\max_{k,i} \tilde{N}_{ki})^2} \right)^{1/2} = O\left( \sqrt{\frac{K}{\max_k \tilde{N}_{kj}} + \frac{JK}{\max_{k,i} \tilde{N}_{ki}}} \right).$$

JG: The idea is then that this can be used to give a regret bound that is order-optimal w.r.t. $T$. Shouldn't be too tricky I don't think as the dependent sampling has been accounted for by this point, and a version of the regret analysis exists for known position biases.

# C Proof of Regret Upper Bound

In this section we provide a proof of Proposition 1, giving an upper bound on the regret of the Epoch-UCB algorithm for the known position bias setting. The proof has two main stages. First we construct an event that all the UCB indices remain in intervals of certain width around the unknown attractiveness parameters, and verify that this is a high-probability event. We simply assume that the regret is the worst possible if this event does not occur. Then conditioned on the high probability event occurring, we utilise bounds on the values of the UCB indices and the properties of the reward function to bound the regret per epoch, which is aggregated over $L$ epochs to give the stated regret bound.

## C.1 Proof of Proposition 1

We begin by defining the regret specifically for an epoch-based algorithm. We have that the $T$-round regret as defined in (2) may be rewritten as

$$Reg(T) = \mathbb{E}\Big(\sum_{l=1}^{L}\sum_{t\in\mathcal{E}_l} R(\mathbf{a}^*) - R(\mathbf{a}^l)\Big) = \mathbb{E}\Big(\sum_{l=1}^{L} |\mathcal{E}_l|\Big(R(\mathbf{a}^*) - R(\mathbf{a}^l)\Big)\Big).$$

We recall that $|\mathcal{E}_l|$, the number of rounds in epoch $l$ follows a Geometric distribution when conditioned on $\mathbf{a}^l$, and that $\mathbf{a}^l$ is a deterministic function of the history $\mathcal{H}_{l-1}$. As such we may use the the law of conditional expectations to replace $|\mathcal{E}_l|$ with its expectation. We have

$$Reg(T) = \mathbb{E}\Bigg(\sum_{l=1}^{L}\mathbb{E}\Big(|\mathcal{E}_l|\Big(R(\mathbf{a}^*) - R(\mathbf{a}^l)\Big) \mid \mathcal{H}_{l-1}\Big)\Bigg) = \mathbb{E}\Big(\sum_{l=1}^{L}\Big(1+\sum_{k=1}^{K}\lambda_k\alpha_k^l\Big)\Big(R(\mathbf{a}^*) - R(\mathbf{a}^l)\Big)\Big).$$

To aid readability, we will define $\Delta R^l = \Big(1+\sum_{k=1}^{K}\lambda_k\alpha_k^l\Big)\Big(R(\mathbf{a}^*) - R(\mathbf{a}^l)\Big)$ for each epoch $l \in [L]$ so that we have

$$Reg(T) = \mathbb{E}\Big(\sum_{l=1}^{L}\Delta R^l\Big).$$

Next, we define a series of events $\mathcal{B}_l, l \in [L]$ which concern the value of the UCBs, as follows,

$$\mathcal{B}_l = \bigcup_{j=1}^{J}\left\{\alpha_{j,l}^{UCB} \notin \left[\alpha_j, \alpha_j + (1+\sqrt{2\alpha_j})\sqrt{\frac{8\log(Jl^2/2)}{\Lambda_{j,l}}} + \frac{4(2+\sqrt{2})\log(Jl^2/2)}{\Lambda_{j,l}}\right]\right\}.$$

We can bound the probability of this event using the concentration results derived in Lemma 3. Specifically, we have that

$$\begin{aligned}
P(\mathcal{B}_l) &= \sum_{j=1}^{J} P(\alpha_{j,l}^{UCB} < \alpha_j) + P\left(\alpha_{j,l}^{UCB} > \alpha_j + (1+\sqrt{2\alpha_j})\sqrt{\frac{8\log(Jl^2/2)}{\Lambda_{j,l}}} + \frac{4(2+\sqrt{2})\log(Jl^2/2)}{\Lambda_{j,l}}\right) \\
&\leq \sum_{j=1}^{J}\frac{3}{Jl} + P\left(|\bar{\alpha}_{j,l}^{UCB} - \alpha_j| > (1+\sqrt{2\alpha_j})\sqrt{\frac{8\log(Jl^2/2)}{\Lambda_{j,l}}} + \frac{4(2+\sqrt{2})\log(Jl^2/2)}{\Lambda_{j,l}}\right) \qquad (41)
\end{aligned}$$

since $P(\alpha_{j,l}^{UCB} < \alpha_j)$ is bounded for each $j \in [J]$ whether $\min(1, 2\bar{\alpha}_{j,l})$ is 1 or $2\bar{\alpha}_{j,l}$. Fixing $j \in [J]$ and considering a single summand from (41) we have,

$$
\begin{aligned}
&P\bigg(|\bar{\alpha}_{j,l}^{UCB} - \bar{\alpha}_j(l)| + |\bar{\alpha}_j(l) - \alpha_j| > (1+\sqrt{2\alpha_j})\sqrt{\frac{8\log(Jl^2/2)}{\Lambda_{j,l}} + \frac{4(2+\sqrt{2})\log(Jl^2/2)}{\Lambda_{j,l}}}\bigg) \\
&\leq P\bigg(|\bar{\alpha}_{j,l}^{UCB} - \bar{\alpha}_j(l)| > \sqrt{\frac{16\alpha_j\log(Jl^2/2)}{\Lambda_{j,l}} + \frac{4(1+\sqrt{2})\log(Jl^2/2)}{\Lambda_{j,l}}}\bigg) \\
&\quad + P\bigg(|\bar{\alpha}_j(l) - \alpha_j| > \sqrt{\frac{8\log(Jl^2/2)}{\Lambda_{j,l}} + \frac{4\log(Jl^2/2)}{\Lambda_{j,l}}}\bigg) \\
&\leq P\bigg(\sqrt{\frac{4\min(1,2\bar{\alpha}_{j,l})\log(Jl^2/2)}{\Lambda_{j,l}} + \frac{4\log(Jl^2/2)}{\Lambda_{j,l}}} > \sqrt{\frac{16\alpha_j\log(Jl^2/2)}{\Lambda_{j,l}}} + (1+\sqrt{2})\frac{4\log(Jl^2/2)}{\Lambda_{j,l}}\bigg) + \frac{4}{Jl} \\
&= P\bigg(\sqrt{\frac{4\min(1,2\bar{\alpha}_{j,l})\log(Jl^2/2)}{\Lambda_{j,l}}} > \sqrt{\frac{16\alpha_j\log(Jl^2/2)}{\Lambda_{j,l}}} + \frac{4\sqrt{2}\log(Jl^2/2)}{\Lambda_{j,l}}\bigg) + \frac{4}{Jl} \\
&\leq P\bigg(\frac{8\bar{\alpha}_{j,l}\log(Jl^2/2)}{\Lambda_{j,l}} > \frac{16\alpha_j\log(Jl^2/2)}{\Lambda_{j,l}} + \frac{32\log^2(Jl^2/2)}{\Lambda_{j,l}^2}\bigg) + \frac{4}{Jl} \\
&= P\bigg(\bar{\alpha}_{j,l} > 2\alpha_j + \frac{4\log(Jl^2/2)}{\Lambda_{j,l}}\bigg) + \frac{4}{Jl} \leq \frac{6}{Jl}. \qquad (42)
\end{aligned}
$$

The first inequality uses the triangle inequality, the second an application of equation (9), the third bounds by replacing the minimum with the $2\bar{\alpha}_{j,l}$ term, and the final uses (11). It follows, combining (41) and (42), that for any $l \in [L]$, $P(\mathcal{B}_l) \leq 9/l$.

Having established $\mathcal{B}_l$ as a low probabilitiy event we will decompose the regret according to $\mathcal{B}_l$, and bound it separately under the events $\mathcal{B}_l$ and $\neg\mathcal{B}_l$. Under $\mathcal{B}_l$ we will resort to trivial bounds on regret coming from the maximum of the reward function, but these will make limited contribution to the overall expected regret, since $\mathcal{B}_l$ occurs with low probability. On $\neg\mathcal{B}_l$, the parameters are bounded in a way that we can exploit to bound the per-round regret with quantities leading to an optimal overall bound. Specifically, we have for $l \in [L]$,

$$
\begin{aligned}
\mathbb{E}\big(\Delta R^l\big) &= \mathbb{E}\Big(\Delta R^l \mathbb{I}\{\mathcal{B}_l\} + \Delta R^l \mathbb{I}\{\neg\mathcal{B}_l\}\Big) \\
&\leq (K+1)P(\mathcal{B}_l) + \mathbb{E}\big(\Delta R^l \mathbb{I}\{\neg\mathcal{B}_l\}\big) \\
&\leq \frac{9(K+1)}{l} + \mathbb{E}\bigg(\Big(1 + \sum_{k=1}^{K}\lambda_k\alpha_k^l\Big)\Big(R(\mathbf{a}^*) - R(\mathbf{a}^l)\Big)\mathbb{I}\{\neg\mathcal{B}_l\}\bigg). \qquad (43)
\end{aligned}
$$

Since the reward function $R$ is monotonically increasing in the attractiveness parameter vector, and under $\neg\mathcal{B}_l$ we have $\alpha_{j,l}^{UCB} \geq \alpha_j$ for all $j \in [J]$, it follows that we also have

$$R(\mathbf{a}, \boldsymbol{\alpha}_l^{UCB}) \geq R(\mathbf{a}, \boldsymbol{\alpha}), \;\; \forall \mathbf{a} \in \mathcal{A},$$

under $\neg\mathcal{B}_l$. We also have by definition of the UCB algorithm that $R(\mathbf{a}^l, \boldsymbol{\alpha}_l^{UCB}) \geq R(\mathbf{a}^*, \bar{\boldsymbol{\alpha}}_l)$, and thus under $\neg\mathcal{B}_l$ we have

$$R(\mathbf{a}^*) - R(\mathbf{a}^l) \leq R(\mathbf{a}^l, \boldsymbol{\alpha}_l^{UCB}) - R(\mathbf{a}^l, \boldsymbol{\alpha})$$

$$
\begin{aligned}
&= \frac{\sum_{k=1}^{K} \lambda_k \alpha^{UCB}_{l,a^l_k}}{1+\sum_{k=1}^{K} \lambda_k \alpha^{UCB}_{l,a^l_k}} - \frac{\sum_{k=1}^{K} \lambda_k \alpha_{a^l_k}}{1+\sum_{k=1}^{K} \lambda_k \alpha_{a^l_k}} \\
&\leq \frac{\sum_{k=1}^{K} \lambda_k (\alpha^{UCB}_{l,a^l_k} - \alpha_{a^l_k})}{1+\sum_{k=1}^{K} \lambda_k \alpha^{UCB}_{l,a^l_k}} \\
&\leq \frac{\sum_{k=1}^{K} \lambda_k (\alpha^{UCB}_{l,a^l_k} - \alpha_{a^l_k})}{1+\sum_{k=1}^{K} \lambda_k \alpha_{a^l_k}} \qquad (44)
\end{aligned}
$$

Thus, combining (43) and (44) we have the following bound on per-round regret,

$$
\mathbb{E}\big(\Delta R^l\big) \leq \frac{9(K+1)}{l} + \mathbb{E}\left(\sum_{k=1}^{K} \lambda_k \left((1+\sqrt{2})\sqrt{\frac{8\log(Jl^2/2)}{\Lambda_{a^l_k,l}}} + \frac{4(2+\sqrt{2})\log(Jl^2/2)}{\Lambda_{a^l_k,l}}\right)\right).
$$

It follows that

$$
\begin{aligned}
Reg(T) &\leq \mathbb{E}\left(\sum_{l=1}^{L}\left[\frac{9(K+1)}{l} + \sum_{k=1}^{K} \lambda_k \left((1+\sqrt{2})\sqrt{\frac{8\log(Jl^2/2)}{\Lambda_{a^l_k,l}}} + \frac{4(2+\sqrt{2})\log(Jl^2/2)}{\Lambda_{a^l_k,l}}\right)\right]\right) \\
&\leq 9(K+1)(\log(T)+1) + \mathbb{E}\left(\sum_{l=1}^{L}\sum_{k=1}^{K}\left(\sqrt{\frac{48\log(Jl^2/2)}{\lambda_K \sum_{s=1}^{l}\mathbb{I}\{a^l_k \in \mathbf{a}_s\}}} + \frac{14\log(Jl^2/2)}{\lambda_K \sum_{s=1}^{l}\mathbb{I}\{a^l_k \in \mathbf{a}_s\}}\right)\right) \\
&\leq 9(K+1)(\log(T)+1) + \frac{14J}{\lambda_K K}\log(JT^2/2)(\log(T)+1) \\
&\qquad + \mathbb{E}\left(\sum_{l=1}^{L}\sum_{k=1}^{K}\sqrt{\frac{48\log(Jl^2/2)}{\lambda_K \sum_{s=1}^{l}\mathbb{I}\{a^l_k \in \mathbf{a}_s\}}}\right) \\
&\leq \left(9(K+1) + \frac{14J}{\lambda_K K}\log(JT^2/2)\right)(\log(T)+1) + \sqrt{\frac{48\log(JT^2)JKT}{\lambda_K}}. \ \square
\end{aligned}
$$

# D Proof of Regret Lower Bound

*Proof:* We first introduce some further notation. For each action $\mathbf{a} \in \mathcal{A}$, a fixed position bias vector $\boldsymbol{\lambda}$, and some constant $\epsilon \in (0, 1/2]$ to be fixed later, define the attractiveness parameter vector $\boldsymbol{\tau}_{\mathbf{a}} \in (0,1]^{J+1}$ such that

$$
\tau_{\mathbf{a},j} = \begin{cases} 1, & j = 0, \\ \frac{1}{S_K} + \frac{\epsilon\lambda_k}{S_{K,2}}, & j = a_k, k \in [K], \\ \frac{1}{S_K}, & \text{otherwise.} \end{cases} \qquad (45)
$$

Let $P_{\mathbf{a}}$ and $\mathbb{E}_{\mathbf{a}}$ denote the law and expectation with respect to the parametrisation $\alpha_j = \tau_{\mathbf{a},j}$. Under $P_{\mathbf{a}}$, the action $\mathbf{a}$ is optimal. We will also make use of additional laws $P_{\mathbf{a}\backslash j'}$ and expectations $\mathbb{E}_{\mathbf{a}\backslash j'}$ for $j' \in \mathbf{a}$, for each $\mathbf{a} \in \mathcal{A}$. Under $P_{\mathbf{a}\backslash j'}$ we set $\alpha_j = \tau_j$, for all $j \neq j'$ and have $\alpha_{j'} = 1/S_K$. Further, for $j \in \mathbf{a}$, introduce the notation $a^{-1}(j)$ to refer to the slot in which action $\mathbf{a}$ places item $j$ - i.e. $a_{a^{-1}(j)} = j$.

For a fixed $\mathbf{a} \in \mathcal{A}$, consider the per-round regret under the problem with parameters $\boldsymbol{\tau}_{\mathbf{a}}$. We have, for any $t \in [T]$,

$$\begin{aligned}
\mathbb{E}_{\mathbf{a}}(r(\mathbf{a}) - r(\mathbf{a}_t)) &= \mathbb{E}_{\mathbf{a}}\left(\frac{1+\epsilon}{2+\epsilon} - \frac{1+\frac{\epsilon}{S_{K,2}}\sum_{k=1}^{K}\lambda_k \sum_{j\in\mathbf{a}}\lambda_{a^{-1}(j)}\mathbb{I}\{a_{k,t}=j\}}{2+\frac{\epsilon}{S_{K,2}}\sum_{k=1}^{K}\lambda_k \sum_{j\in\mathbf{a}}\lambda_{a^{-1}(j)}\mathbb{I}\{a_{k,t}=j\}}\right) \\
&= \mathbb{E}_{\mathbf{a}}\left(\frac{\epsilon - \frac{\epsilon}{S_{K,2}}\sum_{k=1}^{K}\lambda_k \sum_{j\in\mathbf{a}}\lambda_{a^{-1}(j)}\mathbb{I}\{a_{k,t}=j\}}{(2+\epsilon)\left(2+\frac{\epsilon}{S_{K,2}}\sum_{k=1}^{K}\lambda_k \sum_{j\in\mathbf{a}}\lambda_{a^{-1}(j)}\mathbb{I}\{a_{k,t}=j\}\right)}\right) \\
&\geq \mathbb{E}_{\mathbf{a}}\left(\frac{\epsilon - \frac{\epsilon}{S_{K,2}}\sum_{k=1}^{K}\lambda_k \sum_{j\in\mathbf{a}}\lambda_{a^{-1}(j)}\mathbb{I}\{a_{k,t}=j\}}{7}\right).
\end{aligned}$$

It follows that in $T$ rounds, the regret satisfies

$$\begin{aligned}
\max_{\boldsymbol{\alpha}\in(0,1]^J} Reg_{\boldsymbol{\alpha}}(T) &\geq \max_{\mathbf{a}\in\mathcal{A}} \mathbb{E}_{\mathbf{a}}\left(\sum_{t=1}^{T} r(\mathbf{a}) - r(\mathbf{a}_t)\right) \\
&\geq \frac{1}{|\mathcal{A}|}\sum_{\mathbf{a}\in\mathcal{A}} \mathbb{E}_{\mathbf{a}}\left(\sum_{t=1}^{T} r(\mathbf{a}) - r(\mathbf{a}_t)\right) \\
&\geq \frac{1}{7|\mathcal{A}|}\sum_{\mathbf{a}\in\mathcal{A}} \mathbb{E}_{\mathbf{a}}\left(\epsilon T - \frac{\epsilon}{S_{K,2}}\sum_{k=1}^{K}\sum_{j\in\mathbf{a}}\lambda_k\lambda_{a^{-1}(j)}\sum_{t=1}^{T}\mathbb{I}\{a_{k,t}=j\}\right). \qquad (46)
\end{aligned}$$

To further lower bound the RHS expression in (46) we will consider the sum over $t \in [T]$ in isolation.

Define, for $k \in [K], j \in [J]$, and $t \in [T]$, the random variable $N_{k,j}(T) = \sum_{t=1}^{T}\mathbb{I}\{a_{k,t} = j\}$, counting the number of times item $j$ is displayed in slot $k$ over $T$ rounds. For an $\mathbf{a} \in \mathcal{A}$, and $j \in \mathbf{a}$ the expectation of these random variables has the following property,

$$\begin{aligned}
\mathbb{E}_{\mathbf{a}}(N_{k,j}(T)) &\leq \mathbb{E}_{\mathbf{a}\setminus j}(N_{k,j}(T)) + \left|\mathbb{E}_{\mathbf{a}\setminus j}(N_{k,j}(T)) - \mathbb{E}_{\mathbf{a}}(N_{k,j}(T))\right| \\
&\leq \mathbb{E}_{\mathbf{a}\setminus j}(N_{k,j}(T)) + \sum_{s=0}^{T} s\left|P_{\mathbf{a}\setminus j}(N_{k,j}(T)=s) - P_{\mathbf{a}}(N_{k,j}(T)=s)\right| \\
&\leq \mathbb{E}_{\mathbf{a}\setminus j}(N_{k,j}(T)) + T\sum_{s=0}^{T}\left|P_{\mathbf{a}\setminus j}(N_{k,j}(T)=s) - P_{\mathbf{a}}(N_{k,j}(T)=s)\right| \\
&\leq \mathbb{E}_{\mathbf{a}\setminus j}(N_{k,j}(T)) + T||P_{\mathbf{a}\setminus j} - P_{\mathbf{a}}||_{TV} \\
&\leq \mathbb{E}_{\mathbf{a}\setminus j}(N_{k,j}(T)) + T\sqrt{\frac{KL(P_{\mathbf{a}\setminus j} \mid\mid P_{\mathbf{a}})}{2}}.
\end{aligned}$$

Here the final inequality uses Pinsker's inequality. This decomposition is typical of standard regret lower bound analysis.

We now turn our attention to the KL divergence term $KL(P_{\mathbf{a}\setminus j} \mid\mid P_{\mathbf{a}})$. By the Law of Total Entropy (see e.g. Theorem 2.5.3 of Cover and Thomas (2012)), we have

$$KL(P_{\mathbf{a}\setminus j} \mid\mid P_{\mathbf{a}}) = \sum_{t=1}^{T} KL\left(P_{\mathbf{a}}(Q_t \mid Q_1, \ldots, Q_{t-1}) \mid\mid P_{\mathbf{a}\setminus j'}(Q_t \mid Q_1, \ldots, Q_{t-1})\right)$$

$$= \sum_{t=1}^{T} \sum_{\mathbf{a}' \in \mathcal{A}: j \in \mathbf{a}'} \mathbb{I}\{\mathbf{a}_t = \mathbf{a}'\} KL(P_{\mathbf{a}}(Q \mid \mathbf{a}') \mid\mid P_{\mathbf{a}\setminus j'}(Q \mid \mathbf{a}'))$$

$$\leq \sum_{k=1}^{K} N_{k,j}(T) \max_{\mathbf{a}' \in \mathcal{A}: a'_k = j} KL(P_{\mathbf{a}}(Q \mid \mathbf{a}') \mid\mid P_{\mathbf{a}\setminus j}(Q \mid \mathbf{a}')) \tag{47}$$

Here we also use the property that any algorithm gives a deterministic mapping from $Q_{1:t-1}$ to $\mathbf{a}_t$ - since even an instance of a 'randomised' algorithm can, alternatively, be viewed as a deterministic algorithm randomly selected from a (potentially infinitely large) population of algorithms.

The following lemma, whose proof is given in the appendix, bounds the contribution to the KL divergence from a single round.

**Lemma 8** *For two actions $\mathbf{a}, \mathbf{a}' \in \mathcal{A}$, and an item $j \in \mathbf{a}$ such that $a_n = j$ and $a'_{n'} = j$ for some (possibly equal) $n, n' \in [K]$, we have the following bound on the KL divergence between $P_{\mathbf{a}}(Q \mid \mathbf{a}')$ and $P_{\mathbf{a}\setminus j'}(Q \mid \mathbf{a}')$, the marginal distributions on $Q \mid \mathbf{a}'$ implied by the laws $P_{\mathbf{a}}$ and $P_{\mathbf{a}\setminus j}$,*

$$KL\left(P_{\mathbf{a}}(Q \mid \mathbf{a}') \mid\mid P_{\mathbf{a}\setminus j}(Q \mid \mathbf{a}')\right) \leq \frac{17\epsilon^2 \lambda^2_{a'^{-1}(j)} S_K}{S^2_{K,2}}. \tag{48}$$

Thus, combining the decomposition of KL divergence in (47) and the bound on per-round KL divergence in (48) we have for any $j \in \mathbf{a}$,

$$KL(P_{\mathbf{a}} \mid\mid P_{\mathbf{a}\setminus j}) \leq \sum_{k=1}^{K} N_{k,j}(T) \frac{17\epsilon^2 \lambda_k^2 S_K}{S^2_{K,2}}.$$

Moreover, for any $k \in [K]$, $j \in \mathbf{a}$,

$$\mathbb{E}_\tau(N_{k,j}(T)) \leq \mathbb{E}_{\neg j}(N_{k,j}(T)) + T\sqrt{\frac{17\epsilon^2 \lambda_1^2 T S_K}{2 S^2_{K,2}}}.$$

Then combining with (46) we have that the regret is lower bounded, similarly to in Chen and Wang (2018), as,

$$\begin{aligned}
Reg(T) &\geq \frac{1}{7|\mathcal{A}|} \sum_{\mathbf{a} \in \mathcal{A}} \left[ \epsilon T - \frac{\epsilon}{S_{K,2}} \sum_{k=1}^{K} \lambda_k \sum_{j \in \mathbf{a}} \lambda_{a^{-1}(j)} \mathbb{E}_{\mathbf{a}}(N_{k,j}(T)) \right] \\
&\geq \frac{\epsilon T}{7} - \frac{1}{7|\mathcal{A}|} \sum_{\mathbf{a} \in \mathcal{A}} \left[ \frac{\epsilon}{S_{K,2}} \sum_{k=1}^{K} \sum_{j \in \mathbf{a}} \lambda_k \lambda_{a^{-1}(j)} \mathbb{E}_{\mathbf{a}\setminus j}(N_{k,j}(T)) \right] \\
&\qquad - \frac{1}{7|\mathcal{A}|} \sum_{\mathbf{a} \in \mathcal{A}} \left[ \frac{\epsilon^2 T^{3/2}}{S^2_{K,2}} \sum_{k=1}^{K} \sum_{j \in \mathbf{a}} \lambda_k \lambda_{a^{-1}(j)} \sqrt{\frac{17\lambda_1^2 S_K}{2}} \right] \\
&\geq \frac{\epsilon T}{7} - \frac{\epsilon K T}{7J} - \frac{\epsilon^2 T^{3/2}}{7J} \sqrt{\frac{17 S_K}{2 S^2_{K,2}}}
\end{aligned}$$

Finally, we complete the proof by choosing $\epsilon = O(\sqrt{J} S_{K,2} / \sqrt{T S_K})$ and using the assumption that $K \leq J/4$. $\square$

## D.1 Proof of Lemma 8

In this section we provide a proof of Lemma 8, which helps to complete the proof of the regret lower bound, Proposition 2. Lemma 8 gives a bound on the KL divergence between the marginal distributions over a single click variable, under (particular) different attractiveness parameter vectors. The proof makes use of the following result, originally given as Lemma 3 in Chen and Wang (2018), bounding the KL-divergence between categorical random variables.

**Lemma 9 (Lemma 3 of Chen and Wang (2018))** *Suppose $P$ is a categorical distribution on $[M]_0$ with parameters $p_0, \ldots, p_M$, such that if $X \sim P$, $P(X = m) = p_m$ for $m \in [M]_0$. Suppose also that $Q$ is an equivalently defined categorical distribution with parameters $q_0, \ldots, q_M$, and we have $\delta_m = p_m - q_m$ for $m \in [M]_0$. Then*

$$KL(P \mid\mid Q) \leq \sum_{m=0}^{M} \frac{\delta_m^2}{q_m}.$$

*Proof of Lemma 8:* We begin by deriving expressions for the parameters $p_k := P_{\mathbf{a}}(Q = k \mid \mathbf{a}')$,

$$\begin{aligned}
p_0 &= \frac{1}{2 + \epsilon \sum_{l=1}^{K} \lambda_l \sum_{m=1}^{K} \frac{\lambda_m}{S_{K,2}} \mathbb{I}\{a'_l = a_m\}}, \\
p_k &= \frac{\frac{\lambda_k}{S_K} + \epsilon \lambda_k \sum_{m=1}^{K} \frac{\lambda_m}{S_{K,2}} \mathbb{I}\{a'_k = a_m\}}{2 + \epsilon \sum_{l=1}^{K} \lambda_l \sum_{m=1}^{K} \frac{\lambda_m}{S_{K,2}} \mathbb{I}\{a'_l = a_m\}}, \quad k \in [K],
\end{aligned}$$

and $q_k := P_{\mathbf{a}\setminus j}(Q = k \mid \mathbf{a}')$,

$$\begin{aligned}
q_0 &= \frac{1}{2 + \epsilon \sum_{l=1}^{K} \lambda_l \sum_{m=1}^{K} \frac{\lambda_m}{S_{K,2}} \mathbb{I}\{a'_l = a_m, \;\; m \neq n\}}, \\
q_k &= \frac{\frac{\lambda_k}{S_K} + \epsilon \lambda_k \sum_{m=1}^{K} \frac{\lambda_m}{S_{K,2}} \mathbb{I}\{a'_k = a_m, \;\; m \neq n\}}{2 + \epsilon \sum_{l=1}^{K} \lambda_l \sum_{m=1}^{K} \frac{\lambda_m}{S_{K,2}} \mathbb{I}\{a'_l = a_m, \;\; m \neq n\}}, \quad k \in [K].
\end{aligned}$$

To apply Lemma 9 we consider the differences in these parameters. For the no-click event we have

$$p_0 - q_0 = \frac{-\epsilon \lambda_n \lambda_{n'}}{S_{K,2} \left(2 + \epsilon \sum_{l=1}^{K} \lambda_l \sum_{m=1}^{K} \frac{\lambda_m}{S_{K,2}} \mathbb{I}\{a'_l = a_m\}\right) \left(2 + \epsilon \sum_{l=1}^{K} \lambda_l \sum_{m=1}^{K} \frac{\lambda_m}{S_{K,2}} \mathbb{I}\{a'_l = a_m, \;\; m \neq n\}\right)}.$$

For $k \in [K]$ such that $k \neq n'$ we have

$$\begin{aligned}
p_k - q_k &= \frac{\frac{\lambda_k}{S_K} + \epsilon \lambda_k \sum_{m=1}^{K} \frac{\lambda_m}{S_{K,2}} \mathbb{I}\{a'_k = a_m\}}{2 + \epsilon \sum_{l=1}^{K} \lambda_l \sum_{m=1}^{K} \frac{\lambda_m}{S_{K,2}} \mathbb{I}\{a'_l = a_m\}} - \frac{\frac{\lambda_k}{S_K} + \epsilon \lambda_k \sum_{m=1}^{K} \frac{\lambda_m}{S_{K,2}} \mathbb{I}\{a'_k = a_m, \;\; m \neq n\}}{2 + \epsilon \sum_{l=1}^{K} \lambda_l \sum_{m=1}^{K} \frac{\lambda_m}{S_{K,2}} \mathbb{I}\{a'_l = a_m, \;\; m \neq n\}}, \\
&= \frac{-\epsilon \frac{\lambda_k \lambda_n \lambda_{n'}}{S_K S_{K,2}} - \epsilon^2 \frac{\lambda_k^2 \lambda_n \lambda_{n'}}{S_{K,2}^2}}{\left(2 + \epsilon \sum_{l=1}^{K} \lambda_l \sum_{m=1}^{K} \frac{\lambda_j}{S_{K,2}} \mathbb{I}\{a'_l = a_m\}\right) \left(2 + \epsilon \sum_{l=1}^{K} \lambda_l \sum_{m=1}^{K} \frac{\lambda_m}{S_{K,2}} \mathbb{I}\{a'_l = a_m, \;\; m \neq n\}\right)}.
\end{aligned}$$

Finally for the slot $n'$ in which item $j$ is placed we have

$$p_{n'} - q_{n'} = \frac{\frac{\lambda_{n'}}{S_K} + \epsilon\frac{\lambda_n\lambda_{n'}}{S_{K,2}}}{2+\epsilon\sum_{l=1}^K\lambda_l\sum_{m=1}^K\frac{\lambda_m}{S_{K,2}}\mathbb{I}\{a'_l=a_m\}} - \frac{\frac{\lambda_{n'}}{S_K}}{2+\epsilon\sum_{l=1}^K\lambda_l\sum_{m=1}^K\frac{\lambda_m}{S_{K,2}}\mathbb{I}\{a'_l=a_m, m\neq n\}}$$
$$= \frac{\frac{\lambda_n\lambda_{n'}}{S_{K,2}}\Big(2\epsilon+\epsilon^2\sum_{l=1}^K\sum_{m=1}^K\frac{\lambda_l\lambda_m}{S_{K,2}}\mathbb{I}\{a'_l=a_m, m\neq n\}\Big) - \epsilon\frac{\lambda_n\lambda_{n'}^2}{S_K S_{K,2}}}{\Big(2+\epsilon\sum_{l=1}^K\lambda_l\sum_{m=1}^K\frac{\lambda_m}{S_{K,2}}\mathbb{I}\{a'_l=a_m\}\Big)\Big(2+\epsilon\sum_{l=1}^K\lambda_l\sum_{m=1}^K\frac{\lambda_m}{S_{K,2}}\mathbb{I}\{a'_l=a_m,\ \ m\neq n\}\Big)}.$$

It follows, subsequent to some further algebra, that we have

$$\frac{(p_0-q_0)^2}{q_0} \leq \frac{\epsilon^2\lambda_n^2\lambda_{n'}^2}{8S_{K,2}^2},\ \frac{(p_k-q_k)^2}{q_k} \leq \frac{7\epsilon^2\lambda_k\lambda_n\lambda_{n'}^2}{8S_{K,2}^3/S_K},\ \text{and}\ \frac{(p_{n'}-q_{n'})^2}{q_{n'}} \leq \frac{9\epsilon^2\lambda_n\lambda_{n'}^2}{8S_{K,2}^2/S_K}.$$

As such, we have the following by Lemma 9, and the assumption that $S_{K,2}\geq 1$

$$KL(P_{\mathbf{a}}(Q\mid\mathbf{a}')\mid\mid P_{\mathbf{a}\setminus j}(Q\mid\mathbf{a}')) \leq \frac{\epsilon^2\lambda_n^2\lambda_{n'}^2}{8S_{K,2}^2} + \sum_{k=1}^K\frac{7\epsilon^2\lambda_k\lambda_n\lambda_{n'}^2}{8S_{K,2}^3/S_K}\mathbb{I}\{k\neq n'\} + \frac{9\epsilon^2\lambda_n\lambda_{n'}^2}{8S_{K,2}^2/S_K}$$
$$\leq \frac{17\epsilon^2\lambda_{n'}^2 S_K}{S_{K,2}^2}.\quad\square$$

# E Proofs of Technical Lemmas

In this section we provide proofs of the remaining technical lemmas arising in the main text.

## E.1 Proof of Lemma 1

The probability of a no-click event given action $\mathbf{a}^l$ is given as

$$p_0(\mathbf{a}^l) = P(Q_t=0|\mathbf{a}_t=\mathbf{a}^l) = \frac{1}{1+\sum_{k=1}^K\lambda_k\alpha_k^l}.$$

It follows that $n^l = |\mathcal{E}_l|-1$, the number of clicks before the no-click event in epoch $l$ is a Geometric random variable with parameter $p_0(\mathbf{a}^l)$. It follows, that conditioned on $n^l$, each $n_k^l$ count may be viewed as a Binomial random variable,

$$n_k^l|n^l \sim Binom(n^l, p_k),$$

where

$$\tilde{p}_k = \frac{\lambda_k\alpha_k}{\sum_{v=1}^K\lambda_v\alpha_v},$$

is the probability of a click on the item in position $k$, given that there is a click.

The moment generating function of a Binomial random variable is of course well-known, and we therefore have

$$\mathbb{E}_\pi(e^{\theta n_k^l}) = \mathbb{E}_{n^l}\big(\mathbb{E}_\pi(e^{\theta n_k^l}|n^l)\big) = \mathbb{E}_{n^l}\big((\tilde{p}_k e^\theta + 1-\tilde{p}_k)^{n^l}\big).$$

We then consider the result that if $X$ is a Geometric random variable with parameter $p$, then for any $\tau$ such that $\tau(1-p) < 1$ we have $\mathbb{E}(\tau^X) = p/(1-\tau(1-p))$. It follows that, for any $\theta < \log(\frac{\lambda_k\alpha_k^l+1}{\lambda_k\alpha_k^l})$, we have

$$\begin{aligned}\mathbb{E}_\pi(e^{\theta n_k^l}) &= \frac{p_0}{1-(\tilde{p}_k e^{\theta/\lambda_k}+1-\tilde{p}_k)(1-p_0)} \\ &= \frac{p_0}{1-\big(\tilde{p}_k(e^\theta-1)+1\big)(1-p_0)} \\ &= \frac{p_0}{1-(1-p_0)-\lambda_k\alpha_k^l p_0(e^\theta-1)} \\ &= \frac{1}{1-\lambda_k\alpha_k^l(e^\theta-1)},\end{aligned}$$

as stated. We recognise that each $n_k^l \mid \mathbf{a}^l$ is an independent geometric random variable, by considering the moment generating function of the geometric random variable $X$ with parameter $p$ and density $f_X(k) = p(1-p)^k$ for $k \in \{0,1,2,\dots\}$, is given as

$$M(t) = \frac{p}{1-e^t(1-p)} = \frac{1}{1-(\frac{1-p}{p})(e^t-1)},$$

and is defined only for $e^t(1-p) < 1$. □

## E.2 Proof of Lemma 2

We will first demonstrate that the log-likelihood function has at most one stationary point on the parameter space $(0,1]^{J+K-1}$, We have that the log-likelihood of data $\mathbf{N}$ given $\tilde{\mathbf{N}}$ and parameters $\boldsymbol{\alpha}, \boldsymbol{\lambda}$ is

$$\log \mathcal{L}(\mathbf{N}; \tilde{\mathbf{N}}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) = \sum_{k=1}^{K}\sum_{j=1}^{J} N_{kj}\log(\alpha_j\lambda_k) - (N_{kj}+\tilde{N}_{kj})\log(1+\alpha_j\lambda_k). \tag{49}$$

Consdier the partial derivatives of the log-likelihood,

$$\frac{\partial \log \mathcal{L}}{\partial \alpha_j} = \sum_{k=1}^{K}\frac{N_{kj}-\alpha_j\lambda_k\tilde{N}_{kj}}{\alpha_j(1+\alpha_j\lambda_k)}, \quad j \in [J], \quad \frac{\partial \log \mathcal{L}}{\partial \lambda_k} = \sum_{j=1}^{J}\frac{N_{kj}-\alpha_j\lambda_k\tilde{N}_{kj}}{\lambda_k(1+\alpha_j\lambda_k)}, \quad k \in [K]\setminus\{1\}.$$

The solutions of the system of equations $\partial \log \mathcal{L}/\partial\alpha_j = 0, \partial \log \mathcal{L}/\partial\lambda_k = 0$, $j \in [J]$, $k \in [K]\setminus\{1\}$, coincide with the solutions of,

$$\sum_{k=1}^{K}(N_{kj}-\alpha_j\tilde{N}_{kj})\prod_{m\neq k}(1+\alpha_j\lambda_m) = 0, \quad j \in [J] \tag{50}$$

$$\sum_{j=1}^{J}(N_{kj}-\lambda_k\tilde{N}_{kj})\prod_{i\neq j}(1+\alpha_i\lambda_k) = 0, \quad k \in [K]\setminus\{1\}. \tag{51}$$

We will demonstrate that the log-likelihood has at most one stationary point on $(0,1]^{J+K-1}$ by showing that the system of equations given by (50) and (51) has at most one solution on $(0,1]^{J+K-1}$.

For each $j \in [J]$ consider the LHS of equation (50) with all $\lambda_2, \ldots, \lambda_K$ fixed in $(0, 1]$. The result is an $O(\alpha_j^K)$ polynomial, which can be decomposed in to the summation of $k$ $O(\alpha_j^K)$ polynomials, $g_k(\alpha_j) = (N_{kj} - \alpha_j \tilde{N}_{kj}) \prod_{m \neq k}(1 + \alpha_j \lambda_m)$, $k \in [K]$. We have roots of $g_k$ at $\alpha_j = N_{kj}/\tilde{N}_{kj}$ and $\alpha_j = -1/\lambda_m$, $\forall m \neq k$, and notice that $g_k(\alpha_j)$ has negative leading order term when $\alpha_j > 0$. Notice that only one of the solutions is positive, and lies in $(0, 1]$ iff $0 < N_{kj} \leq \tilde{N}_{kj}$.

Since each polynomial $g_k$ has negative leading order term, it follows that $\sum_{k=1}^{K} g_k(\alpha_j) = 0$, i.e. equation (50), also has at most one solution in $(0, 1]$, for fixed $\lambda_2, \ldots, \lambda_K$. An analogous argument applied to equation (51) tells us that there is at most one positive solution value in $(0, 1]$ for each $\lambda_k$, $k \in [K] \setminus \{1\}$, coinciding with all other variables being positive.

This tells us that the system of equations where the partial derivatives are set to zero, has at most one solution in $(0, 1]^{K+J-1}$ and as such the log-likelihood function has at most one stationary point on $(0, 1]^{K+J-1}$. From this we deduce that the log-likelihood is either monotonic on $(0, 1]^{K+J-1}$ or unimodal. We have from Wu (1983) that the EM algorithm will converge to the unique MLE if the log-likelihood is unimodal and continuous, and thus that the EM algorithm, Algorithm 1, will converge to the MLEs. $\square$

# F Independent Product Parameter Model

A perhaps more straightforward approach to the unknown position bias variant of the MNL-LTR problem would be to exploit the closed-form distribution of the estimators $\bar{\gamma}_{j,k}(L) = N_{kj}/\tilde{N}_{kj}$ of product parameters $\gamma_{j,k} = \alpha_j \lambda_k$, $j \in [J]$, $k \in [K]$ and build UCBs around these parameters independently. In this section we argue that this is not an appropriate strategy. Specifically, although such an approach can be shown to eventually learn the optimal action, and indeed have sublinear regret, the amount of exploration it requires is prohibitively large in comparison to our proposed strategy.

## F.1 An MNL-Bandit Approach to MNL-LTR with Unknown Position Biases

We notice that by modelling the $\gamma_{jk}$ parameters as independent, the unknown position bias variant of MNL-LTR can also be thought of as a *constrained MNL-bandit* problem. We can design such a formulation where the decision-maker is oblivious to the ranking aspect, but the constraints on their actions ensure that they implicitly assign a single item to a single slot.

In such a setting there are $JK$ *objects*, indexed $S_{j,k}$ for $j \in [J]$, $k \in [K]$, where selecting object $S_{j,k}$ represents placing item $j$ in slot $k$. Object $S_{j,k}$ is therefore associated with parameter $\gamma_{jk}$. In each round $t \in [T]$ the decision-maker chooses a set $\mathbf{S}_t$ containing $K$ of the objects. The constraints of the associated MNL-LTR problem are such that in each round exactly one of these objects must have index $k$ for each $k \in [K]$.

Introducing $JK$ further indicator variables $x_{jk}(t) = \mathbb{I}\{S_{j,k} \in \mathbf{S}_t\}$ for each $t \in [T]$, these constraints may be expressed as,

$$\sum_{j=1}^{J} x_{j,k}(t) = 1 \ \forall k \in [K], \quad \text{and} \quad \sum_{k=1}^{K} x_{j,k}(t) \leq 1 \ \forall j \in [J], \quad \forall t \in [T]. \tag{52}$$

The first constraint captures the rule that every slot is utilised, and the second constraint captures the rule that each item is used at most once. A valid set of objects $\mathbf{S}_t$, satisfying (52), maps in a one-to-one fashion to a valid action $\mathbf{a}_t \in \mathcal{A}$ for our MNL-LTR problem.

Both Agrawal et al. (2017) and Agrawal et al. (2019) derive order-optimal guarantees for constrained MNL-bandit algorithms. However, the constraint considered in Agrawal et al. (2017) is only a simple cardinality constraint - i.e. an upper limit on the number of objects chosen in each round. Thus the guarantees on the Thompson Sampling approach proposed therein do not carry to the problem with constraints (52). Agrawal et al. (2019), however, allow for a more general class of constraints, requiring that constraints may be expressed in the form $A\mathbf{x} \leq \mathbf{b}$, where $A$ is a totally unimodular (TU) matrix, and $\mathbf{b}$ is an integer-valued vector. The $JK \times (J + 2K)$ matrix $A$ implied by constraints (52) can be shown to be TU. Thus the regret guarantees on the UCB algorithm proposed in Agrawal et al. (2019) apply directly to the constrained MNL-bandit instance associated with an MNL-LTR instance.

Algorithm 4 describes a modification of such an algorithm to incorporate our sharper concentration results. Then, the corollary below gives the corresponding order result on regret enjoyed by this algorithm. The proof of this result is omitted as it follows directly from the observation that the coefficient matrix implied by constraints (52) is TU, and substituting the concentration results of Lemma 3 in to the proof of Theorem 1 of Agrawal et al. (2019).

**Corollary 2** *There exist constants $C_1, C_2 > 0$ such that for any MNL-LTR problem where the item attractiveness parameters satisfy $\alpha_j \leq \alpha_0 = 1$, $j \in [J]$ and the position biases satisfy $\lambda_1 = 1$, $\lambda_k \leq \lambda_1$ $k \geq 2$, the regret in $T$ rounds of Algorithm 4 satisfies*

$$Reg(T) \leq C_1 \sqrt{JKT \log(JKT^2)} + C_2 JK \log^2(JKT).$$

Although this implies a guarantee on the performance of Algorithm 4 which is near optimal in its dependence in $T$, we see that the $O(\log(T))$ term has a worse dependence on $JK$ than, for instance, the Epoch-UCB algorithm for the known position bias case. This hints to the critical issue with deploying Algorithm 4, that its exploration cost scales linearly with the *product* of the number of items and number of slots. In the experiments in Section 6, in particular problem (c), we see that even in 'simple' problems - where the optimal action can be identified quickly - Algorithm 4 is slow to converge.

**Algorithm 4** Epoch-UCB algorithm for MNL-bandit model of MNL-LTR

Initialise with $l = 0$, and $Q_0 = 0$. Iteratively perform the following for $t \in [T]$,

If $Q_{t-1} = 0$

- Set $l \leftarrow l + 1$
- Calculate UCBs. For $j \in [J]$, $k \in [K]$ compute,

$$\gamma_{j,k,l}^{UCB} = \bar{\gamma}_{j,k}(l-1) + \sqrt{\frac{4\min(1, 2\bar{\gamma}_{j,k}(l-1))\log(JKl^2/2)}{\sum_{s=1}^{l-1}\mathbb{I}\{S_{j,k} \in \mathbf{S}_l\}}} + \frac{4\log(JKl^2/2)}{\sum_{s=1}^{l-1}\mathbb{I}\{S_{j,k} \in \mathbf{S}_l\}}.$$

- Solve the optimisation problem for object indicator variables

$$\begin{aligned} \mathbf{x}_l = \underset{x_{j,k}, j\in[J], k\in[K]}{\operatorname{argmax}} & \sum_{j=1}^{J}\sum_{k=1}^{K} x_{j,k}\gamma_{j,k,l}^{UCB} \\ \text{s.t.} \ & \sum_{j=1}^{J} x_{j,k}(t) = 1 \ \forall k \in [K], \\ & \sum_{k=1}^{K} x_{j,k}(t) \leq 1 \ \forall j \in [J] \end{aligned}$$

- Select an action $\mathbf{a}_t \in \mathcal{A}$ associated with $\mathbf{x}_l \in \{0,1\}^{J\times K}$, and observe click variable $Q_t$,

otherwise, set action $\mathbf{a}_t = \mathbf{a}_{t-1}$, and observe click variable $Q_t$.