Ranking with Popularity Bias: User Welfare under Self-Amplification Dynamics

Guy Tennenholtz * Google Research
Nadav Merlis CREST, ENSAE
Martin Mladenov Google Research
Craig Boutilier Google Research

Abstract

While *popularity bias* is recognized to play a role in recommmender (and other ranking-based) systems, detailed analyses of its impact on *user welfare* have largely been lacking. We propose a general mechanism by which item popularity, item quality, and position bias can impact user choice, and how it can negatively impact the collective user utility of various recommender policies. Formulating the problem as a non-stationary contextual bandit, we highlight the importance of exploration, not to eliminate popularity bias, but to mitigate its negative effects. First, naive popularity-biased recommenders are shown to induce linear regret by conflating item quality and popularity. More generally, we show that, even in linear settings, identifiability of item quality may not be possible due to the confounding effects of popularity bias. However, under sufficient variability assumptions, we develop an efficient UCB-style algorithm and prove efficient regret guarantees. We complement our analysis with several simulation studies.

1 Introduction

The study of growth dynamics in multi-agent systems has a long history across many disciplines, and in various domains. These include: the wealth of individuals in a community [Pareto, 1896], the sizes of firms in an economy [Axtell, 2006], and popularity in social/citation networks [Barabási and Albert, 1999]. A prominent finding is that growth is often driven by positive feedback mechanisms—the larger one is, the more likely one is to grow further. Such "rich get richer" phenomena have been explained by a variety of mechanisms, e.g., *preferential attachment* [Simon, 1955, Barabási and Albert, 1999], which often generate power-law size distributions.

Similar phenomena have been observed in *recommender ecosystems*, often described by the umbrella term *popularity bias* [Bellogín et al., 2017]. Such bias is often assumed to arise because *recommender systems (RSs)* recommend popular items more frequently, which in turns increases their consumption by users, thus amplifying their popularity [Abdollahpouri et al., 2019a, Abdollahpouri and Mansoury, 2020]. This can induce or magnify commonly observed long-tail effects. Despite this, compelling and precise mechanisms to explain popularity bias in RSs have yet to be proposed. Specifically, growth dynamics and popularity in recommender ecosystems are complicated by the presence of the RS's ranking algorithm [Baeza-Yates et al., 2005, Zoghi et al., 2017, Wang et al., 2018, Amato et al., 2019]. RS *ranking policies (or rankers)* can both amplify popularity feedback (e.g., the phenomenon of "going viral") as well as mitigate it (e.g., through fairness of exposure [Singh and Joachims, 2018]). We argue that the ranker's role must be incorporated directly into any mechanistic account of popularity bias.

In this work we propose a framework for a more nuanced study of popularity bias in RSs, with an emphasis on its impact on collective user utility (or *social welfare*). We generally assume that utility depends on the quality of selected items, but it may also incorporate popularity, e.g., for social or cultural reasons. We first outline a general mechanism by which an item's popularity can *influence* a

*Correspondence: guytenn@gmail.com

user's choice/consumption behavior in conjunction with other relevant factors, namely, item quality and position bias. We also describe and analyze various RS ranking policies that can *amplify or mitigate the negative welfare effects of popularity bias* under various observability assumptions. In particular, we show that *exploration* must play an essential role in overcoming such effects.

We formalize the problem as a nonstationary contextual bandit and provide regret analyses to show whether, and under what conditions, a ranking policy can optimize long-term expected user utility. As a warmup, we show how an RS ranker that uses popularity as a proxy for user utility can generate harmful popularity bias and exhibit linear regret. More broadly, we then show that an optimal ranking strategy is not achievable in the general case due to the confounding effect of popularity on the estimation of item quality. However, we prove that, under specific variability assumptions, this bias is statistically identifiable. We develop an efficient algorithm, with sublinear regret, that decouples quality from popularity, while exploiting the positive feedback mechanism.

Our contributions are as follows. In Sec. 2 we propose a novel framework that incorporates item quality, item popularity, and rank/position bias in a non-stationary contextual user choice model. In Sec. 3, we analyze the dynamics of popularity bias caused by both rankers and users, showing they are prone to converge to suboptimal item-selection distributions with linear regret. In Sec. 4 we show the ranking problem is generally not identifiable (even under a linear class assumption) in the presence of popularity-biased users, and prove lower bounds. We then show that, under a sufficient diversity assumption, an efficient solution can be developed by formulating a UCB-style algorithm with efficient regret guarantees. Our research represents a significant step in the understanding of the complex dynamics of self-amplification of popularity in RSs and its implications on user welfare, and, for the first time, provides regret guarantees for ranking policies in settings with popularity bias.

2 Problem Definition

We begin by providing our problem formulation, including the introduction of a general *user choice model* that incorporates item quality (or affinity), rank/position bias, and importantly, *popularity bias* in the determination of a user's probability of selecting or consuming a specific recommended item. We later show how popularity bias complicates the computation of optimal RS policies.

Preliminaries. We use lower case bold letters for vectors (e.g., $\boldsymbol{x}$) and upper case bold letters for matrices (e.g., $\boldsymbol{V}$). For $\boldsymbol{u} \in \mathbb{R}^M$ let

$$z_i(\boldsymbol{u}) = \frac{\exp\{u_i\}}{1 + \sum_{j=1}^{M} \exp\{u_j\}}, \quad 1 \le i \le M,$$

and $z_0(\boldsymbol{u}) = 1 - \sum_{i=1}^{M} z_i(\boldsymbol{u})$. Finally, let $\boldsymbol{z}(\boldsymbol{u}) = (z_1(\boldsymbol{u}), \dots, z_M(\boldsymbol{u}))^T$ be the softmax function.

Problem Setup. Let $\mathcal{D}$ be a corpus of *recommendable* items (e.g., articles, music tracks, videos, products) and $\mathcal{U}$ be a set of users. At each time t, a user $u_t \in \mathcal{U}$ is sampled from some distribution $\mathcal{P}_{\mathcal{U}}$ to seek a recommendation. The RS uses a *ranking policy (or ranker)* $\mathcal{R}$ to select M items from $\mathcal{D}$, ranks them in some chosen order, and presents the resulting (ordered) *slate* of items $\boldsymbol{s}_t$ to user u_t, who then selects (or consumes) at most one item from the slate. Here, $\boldsymbol{s}_t = (I_{1,t}, I_{2,t}, \dots, I_{M,t})$, where $I_{i,t}$ is the item at position/rank $i \le M$ at time t. Let $\mathcal{S}$ denote the set of possible slates. Let $c_t \in \{I_{i,t}\}_{i=1}^{M} \cup \{\emptyset\}$ denote the item selected at time t, and $\boldsymbol{c}_{1:t} = (c_1, \dots, c_t)$ the sequence of selections though time t. We use $\mathbb{E}_t$ to denote expectation w.r.t. all histories through time t. Let $n_t(I) = \sum_{k=1}^{t-1} \mathbb{1}\{\text{item } I \text{ was selected at time } k\}$ be the number of selections of item I up to time t (exclusive). Finally, let $\boldsymbol{n}_t(\boldsymbol{s}) = (n_t(I_1), \dots, n_t(I_M))^T$, and $\boldsymbol{n}_t = \boldsymbol{n}_t(\boldsymbol{s}_t)$.

User Choice Model. A *user choice model* determines the probability with which a user u selects a specific item I from a slate $\boldsymbol{s}$. We introduce a choice model that incorporates popularity bias as one of three factors influencing choice. The first factor, *quality*, reflects the inherent value a user derives from an item, and is captured by a *quality bias* function $\mathrm{qb} : \mathcal{U} \times \mathcal{S} \mapsto \mathbb{R}^M$ that, for any user u, specifies the user-specific quality of each item in slate $\boldsymbol{s}$. *Rank bias*, given by a vector $\mathrm{rb} \in \mathbb{R}^M$, determines the user- and item-independent impact of an item's position in a slate on user choice. Finally, the user-dependent *popularity bias* depends on the sequence of selected items (by all users)

up to time t, and is given by $\mathrm{pb}_t : \mathcal{U} \times \mathcal{S} \times \mathcal{D}^{t-1} \mapsto \mathbb{R}^M$. Note that we make no assumptions here about the mechanisms that produce the popularity bias; but we discuss such mechanisms in Sec. 3.

User disposition at time t is given by $\mathrm{disp}_t(u_t, \boldsymbol{s}_t, \boldsymbol{c}_{1:t-1}) = \mathrm{qb}(u_t, \boldsymbol{s}_t) + \mathrm{pb}_t(u_t, \boldsymbol{s}_t, \boldsymbol{c}_{1:t-1}) + \mathrm{rb}$. This combines the bias factors additively (more general models are possible). Finally, we assume the user u_t at time t selects the item at position i with probability

$$z_i(\mathrm{disp}_t(u_t, \boldsymbol{s}_t, \boldsymbol{c}_{1:t-1})), \quad 1 \leq i \leq M; \tag{1}$$

and $z_0(\mathrm{disp}_t(u_t, \boldsymbol{s}_t, \boldsymbol{c}_{1:t-1})) = 1 - \sum_{i=1}^{M} z_i(\mathrm{disp}_t(u_t, \boldsymbol{s}_t, \boldsymbol{c}_{1:t-1}))$ is the probability of no selection.

Modeling Assumptions. We outline several key assumptions used in our analysis below. First, we assume that all of $\mathrm{qb}(u, \boldsymbol{s})_i, \mathrm{pb}(u, \boldsymbol{s}, \boldsymbol{c})_i, \mathrm{rb}_i \in [-1, 1]$, for all $u \in \mathcal{U}, \boldsymbol{s} \in \mathcal{S}, i \leq M$. Next, we assume that the popularity bias of item I is monotonically increasing in the number of selections induced by the RS and bounded. Specifically:

Assumption 1 (Monotonic and Bounded Popularity)**.** *For any* $1 \leq i \leq M, t \in \mathbb{N}, u \in \mathcal{U}, \boldsymbol{s} = (I_1, \ldots, I_M) \in \mathcal{S}$*, and* $\boldsymbol{c} = (c_1, \ldots, c_{t-1}) \in \mathcal{D}^{t-1}$*,*

$$\mathrm{pb}_t(u, \boldsymbol{s}, \boldsymbol{c})_i = \min\left\{\sum_{k=1}^{t-1} \alpha_{k,i}(I_i, c_k),\ \ b_i(u, \boldsymbol{s})\right\},$$

where $\alpha_{k,i} : \mathcal{D} \times \mathcal{D} \mapsto \mathbb{R}_+$ *are unknown, non-negative selection "increment" mappings, and* $b_i : \mathcal{U} \times \mathcal{S} \mapsto \mathbb{R}_+$ *are unknown upper bounds on popularity bias (for* $1 \leq i \leq M, k \in \mathbb{N}$*). Let* $\boldsymbol{b}(u, \boldsymbol{s}) = (b_1(u, \boldsymbol{s}), \ldots, b_M(u, \boldsymbol{s}))^T$*,* $b_{\max} = \max_{1 \leq i \leq M, u \in \mathcal{U}, \boldsymbol{s} \in \mathcal{S}} b_i(u, \boldsymbol{s})$*, and* $\alpha_{\min} = \min_{I \in \mathcal{D}, 1 \leq i \leq M, t \in \mathbb{N}} \alpha_{t,i}(I, I)$*.*

One popularity bias metric is simply the number of selections, i.e., $\mathrm{pb}_t(u, \boldsymbol{s}, \boldsymbol{c})_i \propto n_t(I_i)$, given by stationary increments $\alpha_{k,i}(I, c) = \alpha_0(I) \cdot \mathbb{1}\{I = c\}$. Such bias often arises when the number of selections, views, likes, etc. of items in a slate are shown to the user.[2] In what follows, we assume $b_{\max} < \infty$ and $\alpha_{\min} > 0$.

Similar to logistic bandits [Faury et al., 2020, Abeille et al., 2021, Amani and Thrampoulidis, 2021], we model quality and popularity bias upper bounds using linear functions. Particularly, each user u and slate $\boldsymbol{s}$ are associated with an embedding vector $\boldsymbol{x}_q(u, \boldsymbol{s}) \in \mathbb{R}^{d_q}$ and $\boldsymbol{x}_p(u, \boldsymbol{s}) \in \mathbb{R}^{d_p}$, respectively; and $\mathrm{qb}(u, \boldsymbol{s})_i = \boldsymbol{x}_q^T(u, \boldsymbol{s})\boldsymbol{\theta}_i^*$ and $b_i(u, \boldsymbol{s}) = \boldsymbol{x}_q^T(u, \boldsymbol{s})\boldsymbol{\phi}_i^*$, where $\boldsymbol{\theta}_i^* \in \mathbb{R}^{d_q}, \boldsymbol{\phi}_i^* \in \mathbb{R}^{d_p}$ $(i \leq M)$ are unknown parameter vectors. [3] As we focus on the disentangling of quality and popularity, we assume that rank bias rb is known. Finally, we adopt a standard boundedness assumption from the GLM and multinomial bandit literature:

Assumption 2 (Boundedness)**.** *For any* $1 \leq i \leq M$*,* $\|\boldsymbol{\theta}_i^*\|_2 \leq L_q, \|\boldsymbol{\phi}_i^*\|_2 \leq L_p$*.*

Optimality Criterion. An RS policy (or ranker) $\mathcal{R} : \mathcal{U} \times \mathcal{H}_t \mapsto \mathcal{S}$ determines the item slate shown to a user given an interaction history (here $\mathcal{H}_t$ is the set of all length $t-1$ histories). Once presented, the user makes a selection using the choice model in Eq. (1), and popularity bias is updated per Assumption 1. We measure the *value* of $\mathcal{R}$ using the total expected *user utility* of selected items: $v_t^{\mathcal{R}} = \mathbb{E}_{u_{t+1}, \ldots, u_T \mathcal{R}}\big[\sum_{k \in [t,T]} \sum_{i \leq M} \mathrm{U}(u_k, \boldsymbol{s}_k)_i \cdot z_i(\mathrm{disp}_k(u_k, \boldsymbol{s}_k, \boldsymbol{c}_{1:k-1})) \,|\, u_k = u, \mathrm{pb}_k = \boldsymbol{b}, \boldsymbol{s}_k = \boldsymbol{s}\,\big]$, where $\mathrm{U} : \mathcal{U} \times \mathcal{S} \mapsto \mathbb{R}$ is some measure of user utility. We generally equate user utility with item quality, i.e., $\mathrm{U} \equiv \mathrm{qb}$, though other measures are possible, e.g., $\mathrm{U} \equiv \mathrm{qb} + c \cdot \mathrm{pb}$, for some $c > 0$. Importantly, the RS only observes user selections, not utility (e.g., user-specific item quality), and typically must *attempt to infer the latent utility* U to optimize its policy. Finally, we let $v_t^{\mathcal{R}} = \mathbb{E}_{u \sim P_{\mathcal{U}}}\big[v_t^{\mathcal{R}}(u, 0, S_1^{\mathcal{R}})\big]$ be the expected value of $\mathcal{R}$ when popularity bias is initialized to zero.

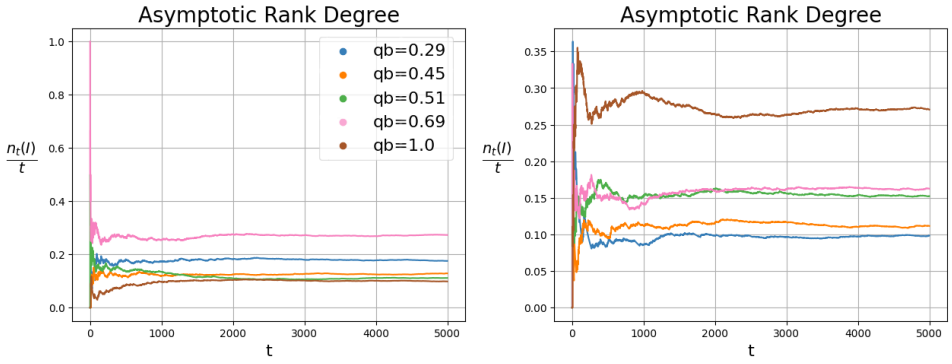
An *optimal ranker* $\mathcal{R}^*$ maximizes the expected return at time t, defined recursively as $\boldsymbol{s}_t^*(u, \boldsymbol{b}) \in \arg\max_{\boldsymbol{s} \in \mathcal{S}} v_t^{\mathcal{R}^*}(u, \boldsymbol{b}, \boldsymbol{s})$. The *cumulative regret* of a ranker $\mathcal{R}$ at time T is then defined by:

$$\mathrm{Reg}_{\mathcal{R}}(T) = v_1^{\mathcal{R}^*} - v_1^{\mathcal{R}}.$$

Our goal is to develop rankers with low regret, where possible. As we will see, this depends critically on the presence, interaction and RS observability of the bias terms that determine user disposition.

[2] We adopt this relatively simple model for clarity, but recognize that, generally, popularity bias may be influenced by *exogenous* factors, and could also be a *stochastic* function of the selection history.

[3] Slate-based features (e.g., $\boldsymbol{x_q}(u, \boldsymbol{s})$) generalize per-item features, but are not necessary in what follows.



(a) Unlucky Draw (b) Lucky Draw

Figure 1: The empirical selection probabilities of different items (of varying qualities, indicated by color) by the popularity-driven ranker as a function of time. (a) An "unlucky" draw where unfavorable effects of the popularity feedback loop are evident: low-quality items I move into higher positions on the slate and remain there, and hence have high asymptotic selection probability $\lim_{t\to\infty} n_t(I)/t$. (b) A "lucky" draw in which high-quality items move to the top of the slate and hence have high asymptotic selection probability.

3 Mechanisms for Popularity Feedback Loops

Before addressing the problem of optimal rankers, we first demonstrate how popularity bias can negatively impact expected value (or social welfare). To illustrate the factors that contribute to popularity bias, we consider three distinct ways in which bias can be induced, including: (i) an RS ranking policy; (ii) via inherent user bias; or (iii) by the behavior of suppliers of items (e.g., product vendors, content creators). In this section, we adopt a simple realization of our framework: we assume N items, with user-item embeddings $\boldsymbol{x}(u, I)$ and parameters $\boldsymbol{\theta}^*$ sampled uniformly on $[0, 1/\sqrt{d}]^d$. This gives a slate embedding $\boldsymbol{x}(u, \boldsymbol{s}) \in \mathbb{R}^{Md}$. In what follows, we show the effect of different modeling choices on ranking dynamics and overall regret, and illustrate the importance of exploration.

Popularity Bias Caused by Rankers. Contemporary RSs often attempt to maximize user utility in a myopic or greedy fashion, causing them to underexplore, and rarely account for certain structural biases, such as popularity (though rank bias is often incorporated). We show that such misspecifications can generate undesirable outcomes and induce linear regret. We first consider the case where users are unbiased by popularity (i.e., $\mathrm{pb} = 0$), only selecting items based on their quality and rank position. For this, we examine the *popularity-driven ranker*, which ranks items at any stage t according to their current popularity (i.e., the number selections so far). This strategy is natural for an RS that cannot observe item quality, but attempts to exploit the correlation between user selections and quality to indirectly estimate item quality using popularity as a proxy.

Fig. 1 depicts the dynamics of the popularity-driven ranker. While user dispositions (hence selections) are positively correlated with quality, due to rank/position bias, the ranker gradually *amplifies* the positions of higher-ranked items given their biased selections. Of course, the ranker may get lucky: if higher quality items are selected more often in early stages of the process, the ranker may converge to high-utility slates. However, Fig. 1 shows that this process may also converge to highly suboptimal rankings—due to the feedback loop between user rank bias in selection and the ranker's own popularity bias—inducing linear regret and low user welfare. We refer the reader to Appendix A for theoretical analysis of these dynamics.

We note that both *estimation of qualities* and *exploration* are vital to ensuring sublinear regret. In fact, if users are unbiased by popularity, an efficient ranker can be constructed using a careful reduction to the multinomial logistic bandits formulation of Amani and Thrampoulidis [2021] (see Appendix B.1 for discussion and proofs).

Popularity-Biased Users. We next turn to situations where feedback loops emerge due to popularity bias in user selection (i.e., $\mathrm{pb} \neq 0$), as defined in Sec. 2. In this setting, a low-quality item might have higher probability of selection than a high-quality item that is less popular. We consider three

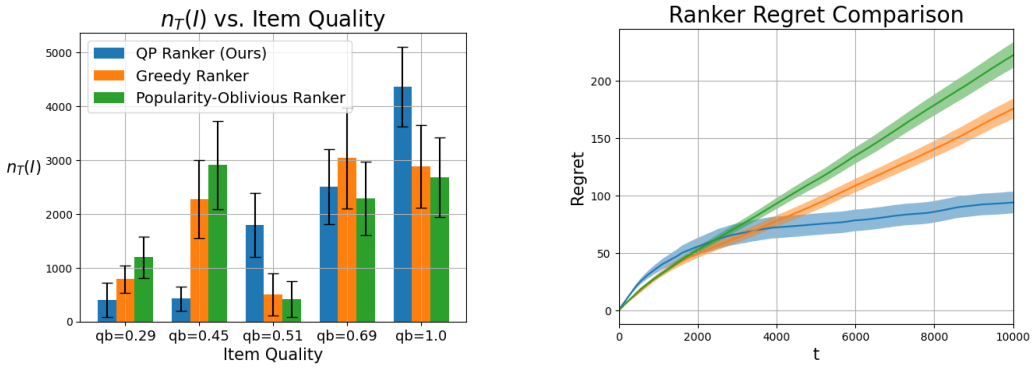


Figure 2: Comparing the QP, greedy and popularity-oblivious rankers on popularity-biased users. Left: number of selections per item (items of different qualities). Right: overall regret. We see that not accounting for user popularity bias increases overall regret significantly; by contrast our QP ranker has low regret and tends to rank high-quality items higher (inducing more high-quality selection).

rankers. The first is an efficient, exploratory ranker, which accounts for popularity bias, which we dub the *Quality-Popularity (QP) Ranker* (its derivation is provided in Sec. 4.4). We compare our QP ranker to a *popularity-oblivious ranker*, which fails to account for user popularity bias, but simply treats user selections as a direct signal to estimate item quality. We implemented this ranker as an efficient exploratory ranker which (incorrectly) assumes $\mathrm{pb} \equiv 0$. Finally, we compare these rankers to a *greedy ranker*, which is identical to our QP-ranker, but with the exploration bonus set to zero.

Fig. 2 compares the QP ranker, the popularity-oblivious ranker and the greedy ranker in terms of regret and item selections. It is evident that the oblivious ranker has significantly worse regret than our QP ranker due to misspecification, and the greedy ranker has worse regret due to its failure to explore efficiently. Fig. 2 also shows the number of selections of each item at the end of training (each is labeled with its quality). The popularity-oblivious ranker tends to rank low quality items higher, inducing more selections of such items via the dynamics of amplification.

Finally, we note that the popularity-oblivious ranker may actually perform reasonably well in the case where user utility is dependent on both item quality *and* popularity, e.g., if $\mathrm{U} \equiv \mathrm{qb} + c \cdot \mathrm{pb}$ for some $c > 0$. Such a case might arise when a user's utility is dictated not only by inherent item quality, but also by positive "network effects" that are correlated with popularity (e.g., the number of friends with whom the user can discuss a recent movie). That said, the oblivious ranker will still not be efficient w.r.t. this combined utility as disambiguation of popularity and quality and explicit exploration are still required. We will delve into the disambiguation of quality from popularity in a later section.

Resource (or Skill) Bias. For completeness, we consider a final mechanism in which popularity bias is implicit in *dynamic item quality*. Consider a model in which users consume new, changing items (e.g., dynamic content) offered by a set of *providers*. In many cases, the popularity of the items offered by a provider influences their *proficiency* at item generation, due to e.g., improved skill or greater resource availability.

Such *resource (or skill) bias* can be modeled by the *increase in quality bias* of a provider's new items I_{new} with the number of selections of its (collective) past items I_{prev}. In other words, $\mathrm{qb}_t(u, I_{\text{new}}) = f_u(n_t(I_{\text{prev}}))$ for some monotonic f_u. A ranker which attempts to maximize user utility greedily recommends to a user the items with highest estimated quality. Nevertheless, quality increments are directly driven by the probabilistic realization of past selections, thus amplifying the increase in quality of the items or providers with more past selections. This, in turn, can induce linear regret, similar to the amplification results above. Moreover, "small" or less popular providers are not given the same chance to improve their quality, as they do not (yet) have the same level of resources or skill.

Resource bias can be directly modeled using the user choice model in Sec. 2, where pb remains a monotonic function of selections, but induces increased resources/skill. Unlike our current model, resource bias directly affects item quality, thereby increasing overall user utility. While this phenomenon is interesting in its own right, we leave its solution to future work.

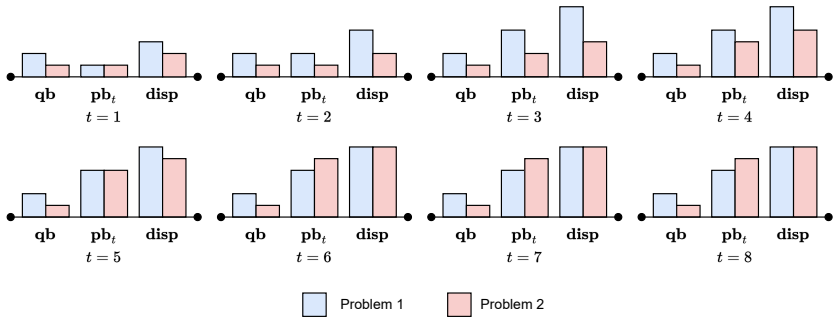


Figure 3: A plot depicting the non-identifiability problem of ranking with popularity bias. As the popularity bias of Problem 1 (blue) and Problem 2 (red) increase, it eventually reaches saturation in both problems. In this example, the rank bias is zero, and the saturation of the two problems brings about the same user disposition $\mathrm{disp} = \mathrm{qb} + \mathrm{pb}$. After iteration $t = 6$, the distribution of selection probability in the two problems is the same, rendering the quality bias not identifiable.

4 Overcoming Popularity Bias

We now turn to the problem of designing optimal ranking policies in the presence of popularity-biased users. We first show that, in the general case, the problem is non-identifiable (i.e., one cannot disambiguate popularity bias from quality), inducing linear-regret lower bounds. This suggests further modeling assumptions are needed to derive efficient algorithms with sub-linear regret. Indeed, using a specific diversity assumption w.r.t. the user population, we derive a UCB-style algorithm that effectively decouples popularity and quality, and prove efficient regret bounds. For the rest of this section we focus on the case of $\mathrm{U} \equiv \mathrm{qb}$.

4.1 Nonidentifiability of Qualities

A natural approach to maximizing value has the ranker estimate qualities (qb). However, as the ranker only observes user selections, doing so requires disentangling the effect of popularity bias on observed behavior from that of quality and rank bias. Unfortunately, this problem is generally non-identifiable, as we demonstrate with a simple counterexample. Consider two one-dimensional problems with $d = 1$, slate size $M = 1$, two items $\mathcal{D} = \{I_1, I_2\}$, with qualities $q(I_1) = 1, q(I_2) = -1$ and no rank bias ($\mathrm{rb} = 0$). The two problems differ only in their item features and saturation bounds: in Problem 1, $\theta^* = \epsilon, b(I_1) = 1 - \epsilon, b(I_2) = 1 + \epsilon$, and in Problem 2, $\theta^* = -\epsilon, b(I_1) = 1 + \epsilon, b(I_2) = 1 = \epsilon$, for some $\epsilon > 0$ (see Fig. 3 for an illustration).

Whenever the ranker presents item I_i and it is selected, its popularity bias increases by at least $\alpha_{\min}$. After $C \frac{b(I_i)}{\alpha_{\min}} \log\left(\frac{2}{\delta}\right)$ selections of I_i, with high probability, $n_t(I_i) \geq \frac{b(I_i)}{\alpha_{\min}}$. When this event occurs at time, say, t_0, popularity bias reaches saturation. Then for any time $t \geq t_0$, $\mathrm{pb}_t(u, I, \boldsymbol{c}) = b(u, I)$, and the selection probability of I_i in both Problems 1 and 2 is identical:

$$z_i(\mathrm{qb}(u, I_i) + \mathrm{pb}_t(u, I_i, \boldsymbol{c})) = z_i(\pm\epsilon + (1 \mp \epsilon)) = e/(1 - e).$$

At this point, w.h.p., the selection distributions in both problems are the same. Hence, the ranker is unable to differentiate them further, resulting in linear expected regret (since qualities are not identifiable). Formally, we have the following result (see proof in Appendix D).

Theorem 1 (Non-Identifiability of the Quality Ranker)**.** *Let Assumptions 1 and 2 hold, and assume $\alpha_{\min} > 0$. Then, for any ranking algorithm $\mathcal{R}$, there exists a ranking problem for which the expected regret is lower bounded by $\mathbb{E}[Reg_{\mathcal{R}}(T)] \geq \Omega(T)$.*

We note that a similar problem can arise even if pb does not saturate, as the softmax function z_i itself reaches saturation with increasing popularity, rendering estimation exponentially hard [Amani and Thrampoulidis, 2021]. In the remainder of this section, we take steps to mitigate the identifiability problem in order to achieve sublinear regret. We being by defining a ranker that is optimal when popularity has reached saturation. Then, motivated by the lower bound in Thm. 1, we show how popularity and quality can be disentangled through a variability assumption on quality-embedding features, and construct an efficient, UCB-style exploration algorithm.

4.2 The Quality Ranker

Learning the optimal ranker $\mathcal{R}^*$ requires good estimation of popularity bias dynamics in order to plan effectively. However, since the Markov process underlying our model is non-recurrent (due to ever-increasing popularity), estimation of these dynamics is generally not possible without additional recurrency assumptions (e.g., the ability to reset the environment). To work around this issue, we first consider a baseline ranking policy which assumes stationarity of the Markov process. The stationary *quality ranker* $\mathcal{R}_q$ recommends slates as follows:

$$\boldsymbol{s}^q(u) \in \arg\max_{\boldsymbol{s}\in\mathcal{S}} \sum_{i=1}^{M} [\mathrm{qb}(u,\boldsymbol{s})]_i \cdot z_i(\mathrm{qb}(u,\boldsymbol{s}) + \boldsymbol{b}(u,\boldsymbol{s}) + \mathrm{rb}).$$

Notice that, when $b_i(u,\boldsymbol{s}) = b_i(u)$, the slate recommended by $\mathcal{R}_q$ is comprised of the M highest-quality items in decreasing order.

The quality ranker is, of course, suboptimal in the general case. However, it is optimal in a counterfactual world where all popularity biases have saturated (hence, in which a steady-state distribution of the Markov chain has been reached). As a consequence, $\mathcal{R}_q$ is in fact *asymptotically optimal*—for large enough T, it achieves constant regret w.r.t. the optimal ranker:

Theorem 2 (Asymptotic Optimality of the Quality Ranker)**.** *Let $\delta \in (0,1)$ and assume $\alpha_{\min} > 0$. For any $T \geq 1$, with probability at least $1-\delta$, the regret of $\mathcal{R}_q$ is*

$$\mathit{Reg}_{\mathcal{R}_q}(T) \leq \mathcal{O}\left(\frac{|\mathcal{D}|\, M b_{\max}}{\alpha_{\min}} \log\left(\frac{|\mathcal{D}|}{\delta}\right)\right).$$

A direct corollary of Thm. 2 is that the regret of any ranker $\mathcal{R}$ can be written as $\mathrm{Reg}_{\mathcal{R}}(T) \leq \mathcal{O}\left(v_1^{\mathcal{R}_q} - v_1^{\mathcal{R}} + \frac{|\mathcal{D}| M b_{\max}}{\alpha_{\min}} \log(\frac{|\mathcal{D}|}{\delta})\right)$. This motivates the analysis of $v_1^{\mathcal{R}_q} - v_1^{\mathcal{R}}$ (the regret w.r.t. $\mathcal{R}_q$), as both are equivalent for large enough T.

4.3 Identifiability Through Variability

We showed in Sec. 4.1 that popularity and quality biases cannot be disentangled in the general case. We now show that, under a condition of sufficient variability induced by the user population, these two factors can be decoupled, a fact we exploit below to design a ranking policy with sublinear regret.

Let $\boldsymbol{\Sigma_{qp}}(u,\boldsymbol{s}) = \boldsymbol{x_q}(u,\boldsymbol{s})\boldsymbol{x_p}^T(u,\boldsymbol{s})$ denote the correlation matrix of $\boldsymbol{x_q}$ and $\boldsymbol{x_p}$. Similarly, we use the notations $\boldsymbol{\Sigma_{qq}}, \boldsymbol{\Sigma_{pp}}$. That is, $\mathbb{E}_u[\boldsymbol{\Sigma_{qp}}(u,\boldsymbol{s})] = \mathbb{E}_u\left[\boldsymbol{x_q}(u,\boldsymbol{s})\boldsymbol{x_p}^T(u,\boldsymbol{s})\right]$ captures the correlation of features in qb and $\boldsymbol{b}$, in expectation over the user population. We make the following assumption to ensure identifiability of quality bias:

Assumption 3. *For all $t \leq T$ and $\boldsymbol{s}_t = \boldsymbol{s}_t(u_t)$, exists $\rho \in (0,1)$, such that*

$$\mathbb{E}_{t-1}\left[\begin{pmatrix} \rho\boldsymbol{\Sigma_{qq}}(u_t,\boldsymbol{s}_t) & \boldsymbol{\Sigma_{qp}}(u_t,\boldsymbol{s}_t) \\ \boldsymbol{\Sigma_{pq}}(u_t,\boldsymbol{s}_t) & \boldsymbol{\Sigma_{pp}}(u_t,\boldsymbol{s}_t) \end{pmatrix}\right] \succeq 0.$$

Let $\rho_{\min} \in (0,1)$ be the smallest such ρ.

Similar assumptions have been made in contextual bandit models [Kannan et al., 2018, Chatterji et al., 2020, Bastani et al., 2021, Papini et al., 2021]. Nevertheless, Assumption 3 is in-fact less demanding than previous assumptions, as it only requires positive-semi-definiteness (see Appendix B.2 for explicit comparison to other assumptions). Also, notice that Assumption 3 holds trivially for $\rho = 1$, yet we require it to hold for some $\rho < 1$. The assumption suffices to disentangle quality from popularity by ensuring enough variability exists in the popularity features. Importantly, this assumption will enable the design of an efficient algorithm which accounts for popularity bias.

4.4 The QP Ranker

We turn to the specification and analysis of our key algorithm, the *Quality-Popularity (QP) Ranker* $\mathcal{R}_{qp}$, presented in Algorithm 1, which exploits the identifiability condition in Assumption 3 to disentangle popularity from quality. Let $\boldsymbol{\psi} \in \mathbb{R}^{d_q+d_p}$ be the concatenation $\boldsymbol{\psi} =$

Algorithm 1 QP Ranker: $\mathcal{R}_{qp}$

require: $\delta \in (0,1), \lambda > 0$
initialize: $\mathcal{H} \leftarrow \emptyset, \boldsymbol{V} \leftarrow \lambda \boldsymbol{I}, \tau_{\min} = \frac{8Mb_{\max}}{\alpha_{\min}} \log(\frac{|\mathcal{D}|}{\delta})$
for $t = 1, 2, \ldots$ **do**
  Observe user $u_t \sim P_{\mathcal{U}}$
  $\hat{\boldsymbol{\psi}}^{\mathrm{ML}} \in \arg\max_{\boldsymbol{\psi}} \mathcal{L}_\lambda(\boldsymbol{\psi}|\mathcal{H})$
  $\hat{\boldsymbol{\psi}} \in \arg\min_{\boldsymbol{\psi}} \left\| g(\boldsymbol{\psi}|\mathcal{H}) - g(\hat{\boldsymbol{\psi}}^{\mathrm{ML}}|\mathcal{H}) \right\|_{\boldsymbol{V}^{-1}}$
  $\boldsymbol{s}_t \in \arg\max_{\boldsymbol{s}\in\mathcal{S}} \sum_{i=1}^{M} \left[\mathrm{qb}_{\hat{\boldsymbol{\theta}}}(u_t, \boldsymbol{s})\right]_i z_i \Big(\mathrm{qb}_{\hat{\boldsymbol{\theta}}}(u_t, \boldsymbol{s}) + \mathrm{pb}_{\hat{\boldsymbol{\phi}}}(u_t, \boldsymbol{s}) + \mathrm{rb}\Big) + \epsilon_{t,\delta}(\boldsymbol{s})$
  Select slate $\boldsymbol{s}_t$ and observe selected item c_t
  if $|\{I \in \boldsymbol{s}_t : n_t(I) \geq \tau_{\min}\}| = M$ **then**
    $\mathcal{H} \leftarrow \mathcal{H} \cup (u_t, \boldsymbol{s}_t, c_t)$
    $\boldsymbol{V} \leftarrow \boldsymbol{V} + \boldsymbol{\Sigma}_x(u_t, \boldsymbol{s}_t)$
  end if
end for

$\left(\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_M^T, \boldsymbol{\phi}_1^T, \ldots, \boldsymbol{\phi}_M^T\right)^T$ of the quality and popularity bias parameters, and $\boldsymbol{x}(u, \boldsymbol{s}) \in \mathbb{R}^{d_q + d_p}$ the concatenation $\boldsymbol{x}(u, \boldsymbol{s}) = \left(\boldsymbol{x}_{\boldsymbol{q}}^T(u, \boldsymbol{s}), \boldsymbol{x}_p^T(u, \boldsymbol{s})\right)^T$ of the quality and popularity embeddings. We can then write $\mathrm{qb}(u, \boldsymbol{s}) + \boldsymbol{b}(u, \boldsymbol{s}) = \boldsymbol{x}^T(u, \boldsymbol{s})\boldsymbol{\psi}$. With the convention that $I_{0,k} = \emptyset$ for all $k \leq T$, define the regularized log-likelihood with regularization parameter $\lambda > 0$ by:

$$\mathcal{L}_\lambda(\boldsymbol{\psi}|\mathcal{H}_t) = \sum_{u_k, \boldsymbol{s}_k, c_k \in \mathcal{H}_t} \sum_{i=0}^{M} \mathbb{1}\{c_k = I_{i,k}\} \log\big(z_i\big(\mathrm{qb}_{\boldsymbol{\theta}}(u_k, \boldsymbol{s}_k) + \mathrm{pb}_{\boldsymbol{\phi}}(u_k, \boldsymbol{s}_k) + \mathrm{rb}\big)\big) - \frac{\lambda}{2} \|\boldsymbol{\psi}\|_2^2 . \quad (2)$$

Let $\hat{\boldsymbol{\psi}}_t^{\mathrm{ML}} \in \arg\max_{\boldsymbol{\psi}} \mathcal{L}_\lambda(\boldsymbol{\psi}|\mathcal{H}_t)$ be the maximum likelihood estimator (MLE) at time t, let $\boldsymbol{V}_t = \lambda \boldsymbol{I} + \sum_{u_k, \boldsymbol{s}_k, c_k \in \mathcal{H}_t} \boldsymbol{\Sigma}_x(u_k, \boldsymbol{s}_k)$ be the design matrix, and define $\boldsymbol{g}(\boldsymbol{\psi}|\mathcal{H}_t) = \boldsymbol{\psi} + \sum_{u_k, \boldsymbol{s}_k, c_k \in \mathcal{H}_t} \boldsymbol{z}\big(\mathrm{qb}_{\boldsymbol{\theta}}(u_k, \boldsymbol{s}_k) + \mathrm{pb}_{\boldsymbol{\phi}}(u_k, \boldsymbol{s}_k) + \mathrm{rb}\big) \otimes \boldsymbol{x}(u_k, \boldsymbol{s}_k)$, the latter being a key quantity related to the gradient. Finally, the projected MLE at time t is: $\hat{\boldsymbol{\psi}} \in \arg\min_{\boldsymbol{\psi}} \left\| g(\boldsymbol{\psi}|\mathcal{H}_t) - g\left(\hat{\boldsymbol{\psi}}_t^{\mathrm{ML}|\mathcal{H}_t}\right) \right\|_{\boldsymbol{V}_t^{-1}}$ (Tighter estimators exist, see Amani and Thrampoulidis [2021] for a derivation of the projected MLE for multinomial logistic bandits).

The QP Ranker proceeds in discrete steps. When a user is sampled at time t, $\mathcal{R}_{qp}$ uses the current MLE $\hat{\boldsymbol{\psi}}$ to construct its slate, using an additive bonus $\epsilon_{t,\delta}(\boldsymbol{s})$. The design matrix $\boldsymbol{V}$ is updated only if the popularities of all items in the slate have reached saturation. Let $\epsilon_{t,\delta}(\boldsymbol{s}) = \left(4\sqrt{M} + \frac{1}{\sqrt{1-\rho_{min}}}\right)\gamma_t(\delta)\sqrt{\mathbb{E}_{t-1}\left[\|\boldsymbol{x}(u_t, \boldsymbol{s})\|_{\boldsymbol{V}_t^{-1}}^2\right]}$, where $\gamma_t(\delta) = 4e^4 M^2 \left(\sqrt{\lambda M} L + 2\sqrt{\log(1/\delta) + Md\log(1 + \frac{t}{\lambda d})}\right)$, and we used the notation $d = d_q + d_p$ and $L = L_q + L_p$. We have the following result (its proof is given in Appendix F).

Theorem 3. *Let Assumptions 1 to 3 hold. Let $\delta \in (0,1)$, and $\lambda = \frac{1}{\sqrt{L}}$. Then with probability at least $1 - \delta$, the regret of the QP Ranker $\mathcal{R}_{qp}$ (Algorithm 1) is upper bounded by*

$$Reg_{\mathcal{R}_{qp}}(T) \leq \mathcal{O}\left(M^{2.5} d\sqrt{T}\left(\sqrt{M} + \frac{1}{\sqrt{1-\rho_{min}}}\right)\log\left(1 + \frac{TL}{d}\right)\sqrt{\log(1/\delta)}\right).$$

Thm. 3 shows that, with effective exploration and proper accounting of popularity bias, sublinear regret can be achieved. Still, we emphasize that Assumption 3 was needed to ensure user utility (quality) can be disentangled from popularity, per our lower bound in Thm. 1. Particularly, we see the effect of $\rho_{\min}$ on the overall regret, acting as a hardness parameter for the problem. We also note that the $\mathcal{O}$-notation in Thm. 3 hides the constant of Thm. 2, which may be large, but does not depend on T. Future research can address this with further assumptions on the dynamics of popularity.

The QP Ranker also performs effectively in simulation, as demonstrated in Fig. 2. We also conducted several additional studies of the algorithm, including varying problem parameters and assessing instances where Assumption 3 does not hold (see Appendix C for these experiments).

5 Related Work

Our work is connected to various lines of research that have investigated popularity bias in RSs, logistic bandit models, bandit models for RSs, and power law distributions in RSs.

Popularity Bias in RSs. A growing body of work deals with popularity bias in recommendation settings [Janssen and PraŁat, 2010, Abdollahpouri, 2019, Abdollahpouri et al., 2019b, Elahi et al., 2021, Abdollahpouri et al., 2021]. Debiasing dynamic recommendations, which can be motivated from both a fairness-of-exposure [Morik et al., 2020] and an estimation-of-quality [Zhu et al., 2021] perspective, explicitly aims to disrupt popularity feedback loops. From a social welfare perspective, this is justifiable if one assumes that item popularity has no explicit value. However, when this is not the case, such policies prevent large communities from forming in the ecosystem, which can in turn limit the creation of content with high production value. Our work focus directly on the social welfare impact itself. Zhao et al. [2022] explicitly model the causal influence of popularity bias, positing that high-quality items should become popular to benefit the ecosystem, a perspective very similar to our. They conclude that disentangling popularity and quality through intervention is a key ingredient in recommendation. Our work takes first steps toward *optimal experimental design* for popularity bias.

Bandits. Our work is related to the growing literature of generalized linear bandits [Filippi et al., 2010], where rewards are sampled from a logistic function with linear features [Abeille et al., 2021, Amani and Thrampoulidis, 2021]. Of particular note is are multinomial bandits [Amani and Thrampoulidis, 2021], which extend the logistic formulation to the M-dimensional case. We build on these results in our user-choice model with popularity and position biases, extending the algorithmic approaches to account for dynamic popularity bias and user utility. In addition, research on the use of bandits for ranking is closely related [Kveton et al., 2015, Zhong et al., 2021, Azizi et al., 2023]. Our model proposes a novel user-choice model, which resembles those used in practical concurrent learning models [Tennenholtz et al., 2022, 2023].

Finally, previous work has attempted to deal with confounding effects in bandits [Bareinboim et al., 2015, Krishnamurthy et al., 2018, Tennenholtz et al., 2021]. In our framework, popularity bias serves as a confounder for item quality/user utility. As demonstrated by Thm. 1, further assumptions are needed in order to de-confound these factors.

Power Laws and Preferential Attachment. There are connections between our model and the classical mechanisms proposed for the emergence of of power laws. If we assume that $qb = 0, rb = 0$ and $pb_t \propto \log(n(I))$, our model is identical to that of preferential attachment, which generates Yule-Simon size distributions [Mitzenmacher, 2004]. Another interesting case is when $qb = 0$, $pb_t = 0$, and $rb \propto \log \frac{1}{r^{1+\alpha}}$, and items are ranked by size. This corresponds to the *prestige ranking model* [Janssen and PraŁat, 2010], where power laws emerge due to the selection distribution.

6 Conclusions, Limitations and Societal Impact

Our work provides insights into the impact of popularity bias on user welfare in RSs and proposes methods to mitigate negative welfare impact. Nevertheless, our framework has several limitations. Primarily, our theoretical analysis assumes that user utility depends solely on selected item quality; however, other factors (e.g., social, cultural) may shape utility. We also note that popularity bias is prevalent in other ranking-based systems (e.g., search engines, social media platforms, news aggregators). Our findings should have implications in these cases, but further research is needed to determine their generalizability.

Our results strongly suggest that popularity bias can negatively impact user welfare, especially in naive RSs that use popularity as a proxy for quality. We show that exploration can help mitigate these negative effects, leading to better long-term user utility (and sub-linear rather than linear regret). Moreover, our work highlights the importance of a deeper understanding of the *mechanisms* underlying popularity bias in ranking-based systems. Given the significant role these play in shaping user opinions and decisions, this understanding is crucial for ensuring fairness, diversity and user well-being. Our work contributes to this understanding and should have implications for the design and deployment of ranking systems.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Himan Abdollahpouri. Popularity bias in ranking and recommendation. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 529–530, 2019.

Himan Abdollahpouri and Masoud Mansoury. Multi-sided exposure bias in recommendation. In *ACM KDD Workshop on Industrial Recommendation Systems 2020*, 2020.

Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. Managing popularity bias in recommender systems with personalized re-ranking. *arXiv preprint arXiv:1901.07555*, 2019a.

Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The unfairness of popularity bias in recommendation. *arXiv preprint arXiv:1907.13286*, 2019b.

Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward Malthouse. User-centered evaluation of popularity bias in recommender systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pages 119–129, 2021.

Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3691–3699. PMLR, 2021.

Sanae Amani and Christos Thrampoulidis. Ucb-based algorithms for multinomial logistic regression bandits. *Advances in Neural Information Processing Systems*, 34:2913–2924, 2021.

Flora Amato, Vincenzo Moscato, Antonio Picariello, and Francesco Piccialli. Sos: a multimedia recommender system for online social networks. *Future generation computer systems*, 93:914–923, 2019.

Robert Axtell. Firm sizes: Facts, formulae, fables and fantasies. Technical report, Center on Social and Economic Dynamics Working Paper, 2006.

Javad Azizi, Ofer Meshi, Masrour Zoghi, and Maryam Karimzadehgan. Overcoming prior misspecification in online learning to rank. In *International Conference on Artificial Intelligence and Statistics*, pages 594–614. PMLR, 2023.

Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. Query recommendation using query logs in search engines. In *Current Trends in Database Technology-EDBT 2004 Workshops: EDBT 2004 Workshops PhD, DataX, PIM, P2P&DB, and ClustWeb, Heraklion, Crete, Greece, March 14-18, 2004. Revised Selected Papers 9*, pages 588–596. Springer, 2005.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286 (5439):509–512, 1999.

Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. *Advances in Neural Information Processing Systems*, 28, 2015.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021.

Alejandro Bellogín, Pablo Castells, and Iván Cantador. Statistical biases in information retrieval metrics for recommender systems. *Information Retrieval Journal*, 20:606–634, 2017.

Niladri Chatterji, Vidya Muthukumar, and Peter Bartlett. Osom: A simultaneously optimal algorithm for multi-armed and linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 1844–1854. PMLR, 2020.

Mehdi Elahi, Danial Khosh Kholgh, Mohammad Sina Kiarostami, Sorush Saghari, Shiva Parsa Rad, and Marko Tkalčič. Investigating the impact of recommender systems on user-based and item-based popularity bias. *Information Processing & Management*, 58(5):102655, 2021.

Richard A Epstein. *The theory of gambling and statistical logic*. Academic Press, 2012.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR, 2020.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, 23, 2010.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

Jeannette Janssen and Paweł PraŁat. Rank-based attachment leads to power law graphs. *SIAM Journal on Discrete Mathematics*, 24(2):420–440, 2010.

Sampath Kannan, Jamie H Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. *Advances in neural information processing systems*, 31, 2018.

Akshay Krishnamurthy, Zhiwei Steven Wu, and Vasilis Syrgkanis. Semiparametric contextual bandits. In *International Conference on Machine Learning*, pages 2776–2785. PMLR, 2018.

Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In *International conference on machine learning*, pages 767–776. PMLR, 2015.

Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, 1(2):226–251, 2004.

Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 429–438, 2020.

Matteo Papini, Andrea Tirinzoni, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Leveraging good representations in linear contextual bandits. In *International Conference on Machine Learning*, pages 8371–8380. PMLR, 2021.

Vilfredo Pareto. Cours d'Économie politique professé a l'université de lausanne (in french), 1896.

Herbert Simon. On a class of skew distribution functions. *Biometrika*, 42(3-4):425–440, 1955.

Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2219–2228, 2018.

Guy Tennenholtz, Uri Shalit, Shie Mannor, and Yonathan Efroni. Bandits with partially observable confounded data. In *Uncertainty in Artificial Intelligence*, pages 430–439. PMLR, 2021.

Guy Tennenholtz, Nadav Merlis, Lior Shani, Shie Mannor, Uri Shalit, Gal Chechik, Assaf Hallak, and Gal Dalal. Reinforcement learning with a terminator. In *Advances in Neural Information Processing Systems*, volume 36, 2022.

Guy Tennenholtz, Nadav Merlis, Lior Shani, Martin Mladenov, and Craig Boutilier. Reinforcement learning with history-dependent dynamic contexts. In *International Conference on Machine Learning*. PMLR, 2023.

Weiqing Wang, Hongzhi Yin, Zi Huang, Qinyong Wang, Xingzhong Du, and Quoc Viet Hung Nguyen. Streaming ranking based recommender systems. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 525–534, 2018.

Zihao Zhao, Jiawei Chen, Sheng Zhou, Xiangnan He, Xuezhi Cao, Fuzheng Zhang, and Wei Wu. Popularity bias is not always evil: Disentangling benign and harmful bias for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–13, 2022. doi: 10.1109/TKDE.2022.3218994.

Zixin Zhong, Wang Chi Chueng, and Vincent YF Tan. Thompson sampling algorithms for cascading bandits. *The Journal of Machine Learning Research*, 22(1):9915–9980, 2021.

Ziwei Zhu, Yun He, Xing Zhao, and James Caverlee. Popularity bias in dynamic recommendation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2439–2449, 2021.

Masrour Zoghi, Tomas Tunys, Mohammad Ghavamzadeh, Branislav Kveton, Csaba Szepesvari, and Zheng Wen. Online learning to rank in stochastic click models. In *International conference on machine learning*, pages 4199–4208. PMLR, 2017.

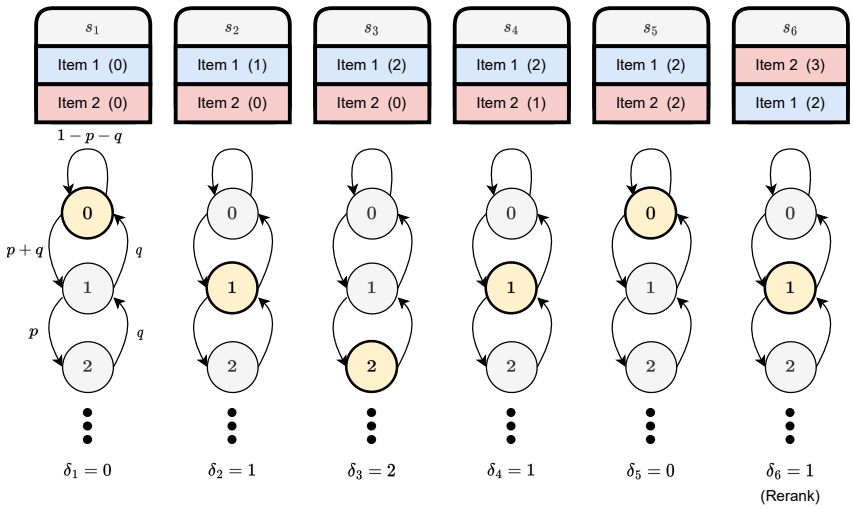


Figure 4: A visualization of the position-biased ranking process for the case of two items. The number of selections on each item is shown in parenthesis. A rerank event occurs whenever the random walk δ_t hits the reflective barrier and exits it through a selection on the second position.

A Dynamics of the Popularity-Driven Ranker

We consider popularity bias that is directly incorporated by suboptimal rankers. For clarity, we focus on the simple case in which users are only biased by position rank (i.e., unbiased by quality or popularity). That is, we assume $\mathrm{qb} \equiv 0, \mathrm{pb} \equiv 0$, such that a user selects position $1 \leq i \leq M$ w.p. $p_i := z_i(\mathrm{rb})$, where $p_1 \geq p_2 \geq \ldots \geq p_M$.

To this end, we wish to anaylze the position dynamics that could emerge, as a ranker accumulates its own popularity bias, and show it may eventually induce a power law distribution over items. For this, we consider a suboptimal, popularity-driven ranker, $\mathcal{R}^{\text{pop}}$, which ranks items based on the number of times they were selected. Specifically, the ranker selects the top M-selected items and ranks them in order of number of selections. Let S_t^{pop} denote the slate induced by the popularity ranker, and let $I_{i,t}$ be the item in position i. Then,

$$I_{i,t} \in \arg \max_{I \in \mathcal{D}, I \notin \{I_{j,t}\}_{j=1}^{i-1}} n_t(I). \qquad \text{(popularity-driven Ranker)}$$

For consistency, we assume the ranker break ties by not reranking any two items if they have the same number of selections. We also assume the initial slate of items is chosen uniformly from the corpus.

The above implies a Markov process, $(\boldsymbol{n}_t)_{t \in \mathbb{N}}$, whose state is defined by the number of selections. In the rest of this section, we will analyze this process to better understand its behavior and the possible emergence of power law distributions. Let $n_t(I_{r,t})$ be the rank degree – a random variable which indicates the number of selections on the item at position r in the ranking at time t. The rank-size distribution $\pi_\infty(r)$ is defined as the fraction of selections of the item at position r at the limit, i.e., $\pi_\infty(r) = \limsup_{t \to \infty} \frac{1}{t} \mathbb{E}[n_t(I_{r,t})]$. Indeed, the rank-size distribution π_∞ is a distribution over rank positions. In what follows, we will show this distribution is monotonically decreasing with rank, whenever users are more inclined to select higher ranked items.

Consider the case of two items, $\mathcal{D} = \{I_1, I_2\}$. We denote the probability of selecting positions 1 and 2 by p and q, respectively. Notice that

$$n_t(I_{r,t}) = \sum_{k=1}^{t-1} \mathbb{1}\{\text{position } r \text{ was selected at time } k\} + \sum_{k=1}^{t-1} \mathbb{1}\{\text{rerank occured at time } k\}. \quad (3)$$

This comes from the fact that the item at position 1 and the item at position 2 can only swap positions if they both have the same size and the item in position 2 was selected. This will lead to the item at position 2 having a larger size and moving to position 1, hence the size at position 1 will increase by 1 even though position 2 was selected.

By Eq. (3), the rank degree is composed of the number of selections on position r, and the total number of reranks. Notice that the expected number of reranks at position r at time t is $p_r t$. To calculate the number of reranks, we denote the random process $\delta_t = n_t(I_{1,t}) - n_t(I_{2,t})$. This quantity evolves according to the following transition kernel

$$\delta_{t+1} = \mathbb{1}\{\delta_t > 0\} \begin{cases} \delta_t + 1 & \text{, w.p. } p \\ \delta_t - 1 & \text{, w.p. } q \\ \delta_t & \text{, w.p. } 1-p-q \end{cases} + \mathbb{1}\{\delta_t = 0\} \begin{cases} 1 & \text{, w.p. } p+q \\ 0 & \text{, w.p. } 1-p-q \end{cases}.$$

Notice that $(\delta_t)_{t\in\mathbb{N}}$ is a homogeneous random walk on the one-dimensional line, with a reflective barrier at 0. A reranking event occurs when, starting at $\delta_1 = 1$, the random walk hits the reflective barrier and then exits it through a selection on the second position (see Fig. 4). That is,

$$P(\text{rerank at time } k > t | \delta_t = 1) = \underbrace{P(\exists k > t, \delta_k = 0 | \delta_t = 1)}_{\text{hit reflective barrier if stated at } \delta_t = 1} \cdot \underbrace{q}_{\text{selection on second position}}. \tag{4}$$

Once a rerank occurs, the random walk restarts at 1 and proceeds identically.

It remains to characterize $P(\exists k > t, \delta_k = 0 | \delta_t = 1)$ in Eq. (4). From the Gambler's Ruin analysis [Epstein, 2012], we have that

$$\begin{aligned} P(\exists k > t, \delta_k = 0 | \delta_t = 1) &= \lim_{N\to\infty} P(\delta_t \text{ hits 0 before } N | \delta_0 = 1) \\ &= \lim_{N\to\infty} \begin{cases} \frac{\left(\frac{q}{p}\right) - \left(\frac{q}{p}\right)^N}{1 - \left(\frac{q}{p}\right)^N} & , p \neq q \\ 1 - \frac{1}{N} & , p = q \end{cases} = \begin{cases} \frac{q}{p} & , p > q \\ 1 & , p = q \end{cases}, \end{aligned}$$

where in the last equality we used the fact that $p \geq q$. It then follows from Eq. (4) that the expected number of reranks at time t is upper bounded by $\frac{1}{1-\frac{q^2}{p}} = \frac{p}{p-q^2}$, for $p > q$. Finally, using Eq. (3), we conclude that, for $p > q$

$$\pi^1_\infty = \lim_{t\to\infty} \frac{1}{t} \mathbb{E}[\text{number of selections on p. } 1 \ + \ \text{number of reranks}] = \lim_{t\to\infty} \frac{1}{t} (p \cdot t + \text{Const}) = p.$$

Similarly, $\pi^2_\infty = q$. This, in turn, means that the dynamic ranking process would eventually freeze on a particular ranking for which one item will always be ranked above another. In other words, as long as the number of reranks is finite, the process would eventually freeze. By symmetry of the process, the ranking will freeze on any slate $\boldsymbol{s} \in \mathcal{S}$ and any order of items with equal probability. Notably, when $p = q$ the number of reranks is infinite, and $\pi^1_\infty, \pi^2_\infty$ diverge, as both items are reranked infinitely often.

B Discussion

B.1 The Case of No Popularity Bias

When no popularity bias exists, our model reduces to the multinomial model of Amani and Thrampoulidis [2021]. We note that Amani and Thrampoulidis [2021] uses tighter confidence sets for the parameters due to exponential dependence on coefficients relating to the gradient of the softmax function $\boldsymbol{z}$. For clarity and simplicity, we chose to bound $\mathrm{qb}, \mathrm{pb}, \mathrm{rb}$ in the interval $[0, 1]$ as to avoid this dependence. Nevertheless, our results easily extend to general intervals, for which case one can apply the tighter parameter estimation guarantees of Amani and Thrampoulidis [2021] to reduce the exponential dependence (yet not eliminate it). While the proofs and derivations of our results would not change, a different estimator would need to be used, projecting to a more involved set.

B.2 Assumption 3

To better understand Assumption 3 we first note the limit case of $\rho = 1$. While the assumptions requires $\rho < 1$ (as also evident by our regret bound in Thm. 3), the assumption always holds for $\rho = 1$, since it is reduced to $\mathbb{E}_{t-1}\left[\begin{pmatrix} \boldsymbol{\Sigma_{qq}}(u_t, \boldsymbol{s}_t) & \boldsymbol{\Sigma_{qp}}(u_t, \boldsymbol{s}_t) \\ \boldsymbol{\Sigma_{pq}}(u_t, \boldsymbol{s}_t) & \boldsymbol{\Sigma_{pp}}(u_t, \boldsymbol{s}_t) \end{pmatrix}\right] \succeq 0$, which holds by definition. It follows that a sufficient condition for Assumption 3 to hold for some $\rho < 1$ is:

$$\mathbb{E}_{t-1}\left[\begin{pmatrix} \boldsymbol{\Sigma_{qq}}(u_t, \boldsymbol{s}_t) & \boldsymbol{\Sigma_{qp}}(u_t, \boldsymbol{s}_t) \\ \boldsymbol{\Sigma_{pq}}(u_t, \boldsymbol{s}_t) & \boldsymbol{\Sigma_{pp}}(u_t, \boldsymbol{s}_t) \end{pmatrix}\right] \succ 0. \tag{5}$$

The assumption in Eq. (5) has been used in previous work on contextual bandits (e.g., [Chatterji et al., 2020]). Nevertheless, we emphasize that Assumption 3 can hold even in cases where Eq. (5) does not hold. For example, consider a case where popularity bias does not exist. In this case, we have

$$\mathbb{E}_{t-1}\left[\begin{pmatrix} \boldsymbol{\Sigma_{qq}}(u_t, \boldsymbol{s}_t) & \boldsymbol{\Sigma_{qp}}(u_t, \boldsymbol{s}_t) \\ \boldsymbol{\Sigma_{pq}}(u_t, \boldsymbol{s}_t) & \boldsymbol{\Sigma_{pp}}(u_t, \boldsymbol{s}_t) \end{pmatrix}\right] = \mathbb{E}_{t-1}\left[\begin{pmatrix} \boldsymbol{\Sigma_{qq}}(u_t, \boldsymbol{s}_t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\right].$$

While clearly $\mathbb{E}_{t-1}\left[\begin{pmatrix} \boldsymbol{\Sigma_{qq}}(u_t, \boldsymbol{s}_t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\right] \not\succ 0$, one can see that $\mathbb{E}_{t-1}\left[\begin{pmatrix} \rho\boldsymbol{\Sigma_{qq}}(u_t, \boldsymbol{s}_t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\right] \succeq 0$, for all $\rho \in (0, 1)$. Indeed, Assumption 3 is less demanding than previous assumptions used in the contextual bandit literature. Moreover, the assumption is used only to disambiguate popularity from quality. To gain intuition as to why Assumption 3 becomes stronger as $\rho \to 0$, notice it does not hold whenever $\rho = 0$ and $\mathbb{E}_{t-1}[\boldsymbol{\Sigma_{qq}}(u_t, \boldsymbol{s}_t)], \mathbb{E}_{t-1}[\boldsymbol{\Sigma_{pp}}(u_t, \boldsymbol{s}_t)] \neq \mathbf{0}$, since the matrix

$$\mathbb{E}_{t-1}\left[\begin{pmatrix} \mathbf{0} & \boldsymbol{\Sigma_{qp}}(u_t, \boldsymbol{s}_t) \\ \boldsymbol{\Sigma_{pq}}(u_t, \boldsymbol{s}_t) & \boldsymbol{\Sigma_{pp}}(u_t, \boldsymbol{s}_t) \end{pmatrix}\right]$$

can be positive semidefinite if and only if $\boldsymbol{\Sigma_{qp}}(u_t, \boldsymbol{s}_t) = \mathbf{0}$. Since we assumed $\mathbb{E}_{t-1}[\boldsymbol{\Sigma_{qq}}(u_t, \boldsymbol{s}_t)], \mathbb{E}_{t-1}[\boldsymbol{\Sigma_{pp}}(u_t, \boldsymbol{s}_t)] \neq \mathbf{0}$, this cannot hold, rendering the assumption empty for $\rho = 0$. Indeed, the regret bound in Thm. 3 improves as ρ nears zero, due to the assumption injecting stronger disambiguation between popularity and quality.

Finally, we show an example for which Assumption 3 does not hold. In such cases disambiguation between quality and popularity is not possible, as shown by our regret lower bound in Thm. 1. Consider, for example a case where

$$\mathbb{E}_{t-1}\left[\begin{pmatrix} \boldsymbol{\Sigma_{qq}}(u_t, \boldsymbol{s}_t) & \boldsymbol{\Sigma_{qp}}(u_t, \boldsymbol{s}_t) \\ \boldsymbol{\Sigma_{pq}}(u_t, \boldsymbol{s}_t) & \boldsymbol{\Sigma_{pp}}(u_t, \boldsymbol{s}_t) \end{pmatrix}\right] = \begin{pmatrix} \mathbf{1}_{d_q \times d_q} & \mathbf{1}_{d_q \times d_p} \\ \mathbf{1}_{d_p \times d_q} & \mathbf{1}_{d_p \times d_p} \end{pmatrix} = \mathbf{1}_{d \times d}.$$

Clearly, $\mathbf{1}_{d \times d}$ is a rank-one matrix, and is positive-semidefinite. Considering Assumption 3, letting $\boldsymbol{x} = (1, 0, 0, \ldots, 0, -1)^T$, we have that

$$\boldsymbol{x}^T \begin{pmatrix} \rho\mathbf{1}_{d_q \times d_q} & \mathbf{1}_{d_q \times d_p} \\ \mathbf{1}_{d_p \times d_q} & \mathbf{1}_{d_p \times d_p} \end{pmatrix} \boldsymbol{x} = \rho - 1 < 0, \forall \rho \in (0, 1).$$

Hence, Assumption 3 does not hold for any $\rho \in (0, 1)$.

To conclude, we believe Assumption 3 to be a reasonable assumption for most use cases. It is less demanding than previous work and captures the hardness of our problem in terms of disambiguation between quality and popularity, as captured by our regret bound in Thm. 3.

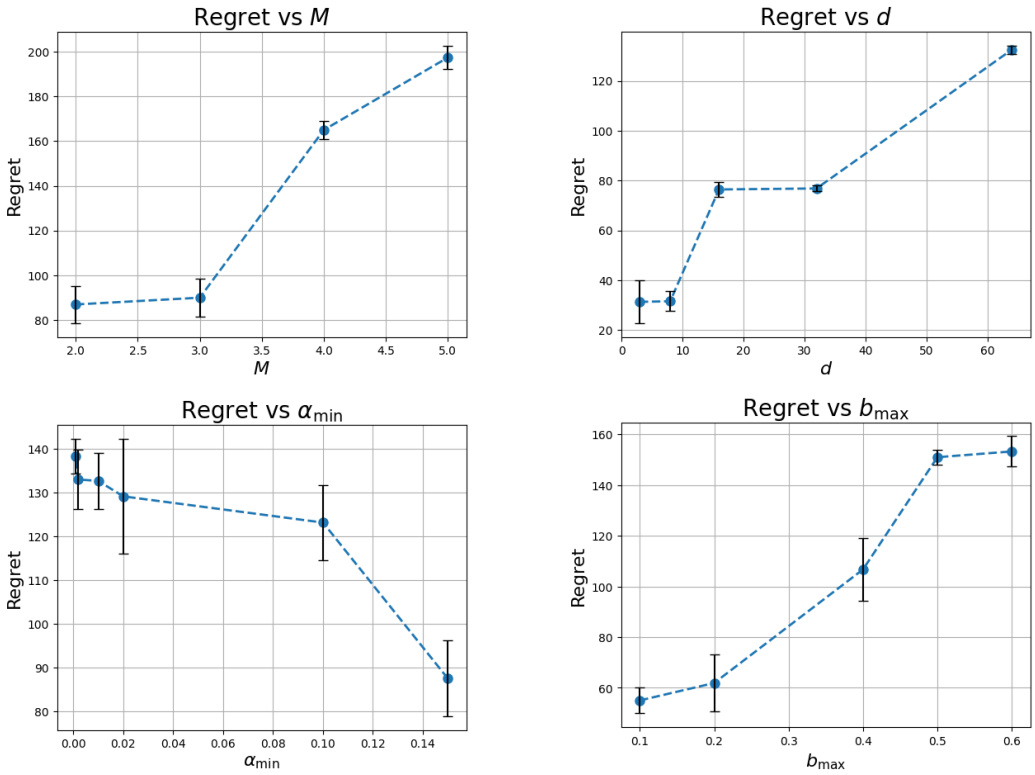


Figure 5: Varying parameters of the QP-Ranker. Results show average regret over final 100 time steps, averaged over 10 seeds. Our results exhibit behavior consistent with our regret.

C Ablations

We tested the QP-Ranker in Algorithm 1 on a series of synthetic environments. For all our experiments we uniformly sampled parameters and embeddings in $\left[0, \frac{1}{\sqrt{d}}\right]^d$. We also scaled the popularity bias, pb w.r.t. $b_{\max}$ to understand the affect of this scaling. We used the same dimension for d_q and d_p, and denote both of them as d here. We varied over problem parameters including $M, d, \alpha_{\min}$, and $b_{\max}$. For each experiment we fixed all parameters to default values, as shown in the table below.

M	d	$\alpha_{\min}$	$b_{\max}$	T
3	8	0.02	0.2	10000

Fig. 5 depicts regret of the QP-Ranker w.r.t. variations in $M, d, \alpha_{\min}$, and $b_{\max}$. While Thm. 3 shows a M^3 dependence in the regret, we found that the dependence to be much better in our synthetic environments. Additionally, we found that, while $\alpha_{\min}$ did significantly increased overall regret (due to its effect in the constant in Thm. 2), we found this effect to diminish. We believe this is due to the ranker not needing to explore the full corpus, for which case saturation of a few items is enough to achieve the desired result. Finally, we found $b_{\max}$ to strongly affect overall regret. This is expected, as lower values of $b_{\max}$ result in a lower upper bound on disp (values lower than 1).

D Proof of Thm. 1

Proof. Without loss of generality, consider the case of a slate containing a single item ($M = 1$) and a corpus of two items $\mathcal{D} = \{I_1, I_2\}$. We assume the problem can be one of two instances; the first instance ν_1 with quality biases $(\mathrm{qb}^1(I_1), \mathrm{qb}^1(I_1)) = (\epsilon, -\epsilon)$, for some small $\epsilon \in (0, 1]$, and the second instance ν_2 with $(\mathrm{qb}^1(I_1), \mathrm{qb}^1(I_1)) = (-\epsilon, \epsilon)$. At every time an item is played, its popularity bias increases by at least $\alpha_{\min} \in (0, 1]$, and the maximal popularity biases are $(b(I_1), b(I_2)) = (1 - \epsilon, 1 + \epsilon)$ in instance ν_1 and $(b(I_1), b(I_2)) = (1 + \epsilon, 1 - \epsilon)$ in instance ν_2. Notice that in both problems, when an item reaches its maximal popularity bias, we have $\mathrm{qb}(I) + \mathrm{pb}(I) = 1 = \mathrm{disp}_{\max}$.

We denote by $n_t^p(I)$, the number of times item I was played up to time $t - 1$, and by $n_t(I)$, the number of time it was selected –generated a reward of 1. We also denote the probability that an item I was selected at round t under instance i by $\mu_t^i(I) = \frac{\exp(\mathrm{qb}^i(I) + \mathrm{pb}_t^i(I))}{1 + \exp(\mathrm{qb}^i(I) + \mathrm{pb}_t^i(I))}$.

Now, recall that if $x \in [-1, 1]$ and $y \in [0, 1 - x]$, using the monotonicity and Lipchitz constant of the logistic function, one can bound

$$\frac{\exp(x + y)}{1 + \exp(x + y)} \geq \frac{\exp(x)}{1 + \exp(x)} + \frac{e}{(1 + e)^2} y \geq \frac{\exp(x)}{1 + \exp(x)} \geq 0.25 + \frac{y}{6}$$

In particular, given an item I of quality $q \in [-1, 1]$, its selection probability lies in the interval

$$\mu_t(I) \in \left[\frac{\exp(q + \alpha_{\min} n_t(I))}{1 + \exp(q + \alpha_{\min} n_t(I))}, \frac{\exp(\mathrm{disp}_{\max})}{1 + \exp(\mathrm{disp}_{\max})} \right] \subseteq \left[0.25 + \frac{\alpha_{\min}}{6} n_t(I), 0.75 \right],$$

where by convention, if the interval is empty then $\mu_t(I) \triangleq \mu^s = 0.75$. Similarly, we denote the distribution of the item click rate in problem i at round t by $\nu_{i,t}$

Assume any fixed strategy and let H_t denote the history of the decision process up to time t, including all internal randomization up to time $t + 1$. Specifically, we denote the selection outcome at time t by Y_t (namely, $Y_t = 1$ if the item was selected) and by U_t, the internal randomization for time $t + 1$. Finally, let $\mathbb{P}_\nu$ denote the probability measure w.r.t. arm distribution ν. Following the derivation of inequality (6) in [Garivier et al., 2019], we have by the chain rule of KL divergence that

$$\begin{aligned} KL(\mathbb{P}_{\nu_1}^{H_{t+1}}, \mathbb{P}_{\nu_2}^{H_{t+1}}) &= KL(\mathbb{P}_{\nu_1}^{(H_t, Y_{t+1}, U_{t+1})}, \mathbb{P}_{\nu_2}^{(H_t, Y_{t+1}, U_{t+1})}) \\ &= KL(\mathbb{P}_{\nu_1}^{H_t}, \mathbb{P}_{\nu_2}^{H_t}) + KL(\mathbb{P}_{\nu_1}^{(Y_{t+1}, U_{t+1})|H_t}, \mathbb{P}_{\nu_2}^{(Y_{t+1}, U_{t+1})|H_t}) \end{aligned}$$

Next, we can write

$$\begin{aligned} KL(\mathbb{P}_{\nu_1}^{(Y_{t+1}, U_{t+1})|H_t}, \mathbb{P}_{\nu_2}^{(Y_{t+1}, U_{t+1})|H_t}) &= \mathbb{E}_{\nu_1}[\mathbb{E}_{\nu_1}[KL(\nu_{1,t}(I_t) \otimes U_{t+1}, \nu_{2,t}(I_t) \otimes U_{t+1}|H_t)]] \\ &= \mathbb{E}_{\nu_1}[\mathbb{E}_{\nu_1}[KL(\nu_{1,t}(I_t), \nu_{2,t}(I_t)|H_t)]] \\ &= \mathbb{E}_{\nu_1}\left[\mathbb{1}\{I_t = I_1\} kl\big(\mu_t^1(I_1), \mu_t^2(I_1)\big) + \mathbb{1}\{I_t = I_2\} kl\big(\mu_t^1(I_2), \mu_t^2(I_2)\big)\right], \end{aligned}$$

where $kl(\mu_1, \mu_2)$ is the KL divergence between two Bernoulli random variables of means μ_1, μ_2. Importantly, notice that the expectation follows the history given problem instance ν_1, thus, if under this instance $n_t(I) > \frac{3}{\alpha_{\min}}$, we will have $\mu_t^1(I) = \mu_t^2(I) = \mu^s$ and $kl(\mu_t^1(I), \mu_t^2(I)) = 0$. Otherwise, since arm means are bounded in $[0.25, 0.75]$, one can easily see that

$$kl(\mu_t^1(I), \mu_t^2(I)) \leq \max\{kl(0.25, 0.75), kl(0.75, 0.25)\} \leq 1$$

Substituting back yields the bound

$$\begin{aligned} &KL(\mathbb{P}_{\nu_1}^{(Y_{t+1}, U_{t+1})|H_t}, \mathbb{P}_{\nu_2}^{(Y_{t+1}, U_{t+1})|H_t}) \\ &\quad \leq \mathbb{E}_{\nu_1}\left[\mathbb{1}\left\{I_t = I_1, n_t(I_1) \leq \frac{3}{\alpha_{\min}}\right\} + \mathbb{1}\left\{I_t = I_2, n_t(I_2) \leq \frac{3}{\alpha_{\min}}\right\}\right], \end{aligned}$$

and iterating over the chain rule of the KL divergence yields

$$KL(\mathbb{P}_{\nu_1}^{H_{t+1}}, \mathbb{P}_{\nu_2}^{H_{t+1}}) \leq \mathbb{E}_{\nu_1}\left[\sum_{s=1}^{t} \mathbb{1}\left\{I_t = I_1, n_t(I_1) \leq \frac{3}{\alpha_{\min}}\right\}\right] + \mathbb{E}_{\nu_1}\left[\mathbb{1}\left\{I_t = I_2, n_t(I_2) \leq \frac{3}{\alpha_{\min}}\right\}\right].$$

We move forward and bound each of these sums. In particular, one can write

$$\begin{aligned}\mathbb{E}_{\nu_1}\left[\mathbb{1}\left\{I_t = I, n_t(I) \leq \frac{3}{\alpha_{\min}}\right\}\right] &= \mathbb{E}_{\nu_1}\left[\mathbb{1}\left\{Y_{t+1} = 1, I_t = I, n_t(I) \leq \frac{3}{\alpha_{\min}}\right\}\right] \\ &\quad + \mathbb{E}_{\nu_1}\left[\mathbb{1}\left\{Y_{t+1} = 0, I_t = I, n_t(I) \leq \frac{3}{\alpha_{\min}}\right\}\right] \\ &\leq \mathbb{E}_{\nu_1}\left[\mathbb{1}\left\{Y_t = 1, I_t = I, n_t(I) \leq \frac{3}{\alpha_{\min}}\right\}\right] \\ &\quad + 0.75\mathbb{E}_{\nu_1}\left[\mathbb{1}\left\{I_t = I, n_t(I) \leq \frac{3}{\alpha_{\min}}\right\}\right].\end{aligned}$$

Reorganizing and noticing that $Y_t = 1$ implies that $n^c_{t+1}(I) = n_t(I) + 1$, we get

$$\mathbb{E}_{\nu_1}\left[\mathbb{1}\left\{I_t = I, n_t(I) \leq \frac{3}{\alpha_{\min}}\right\}\right] \leq 4\mathbb{E}_{\nu_1}\left[\mathbb{1}\left\{n^c_{t+1}(I) = n_t(I) + 1, n_t(I) \leq \frac{3}{\alpha_{\min}}\right\}\right],$$

and summing while noting that both events in the indicator can be active only for $\frac{3}{\alpha_{\min}} + 1$ times yields

$$KL(\mathbb{P}^{H_{t+1}}_{\nu_1}, \mathbb{P}^{H_{t+1}}_{\nu_2}) \leq 8\left(\frac{3}{\alpha_{\min}} + 1\right) \leq \frac{25}{\alpha_{\min}}.$$

The next step of the proof is to apply the data-processing inequality for the KL divergence (Lemma 1 of Garivier et al. 2019), with the random variables $Z = \frac{n_{t+1}(I)}{t} \in [0, 1]$ for any $I \in \{I_1, I_2\}$. By doing so, one get that

$$kl\left(\frac{\mathbb{E}_{\nu_1}[n^p_{t+1}(I)]}{t}, \frac{\mathbb{E}_{\nu_2}[n^p_{t+1}(I)]}{t}\right) \leq KL(\mathbb{P}^{H_{t+1}}_{\nu_1}, \mathbb{P}^{H_{t+1}}_{\nu_2}) \leq \frac{25}{\alpha_{\min}}. \tag{6}$$

Finally, recall that for any $p, q \in (0, 1)$, we have

$$\begin{aligned}kl(p, q) &= p \ln \frac{p}{q} + (1 - p) \ln \frac{1 - p}{1 - q} \\ &= p \ln \frac{1}{q} + \underbrace{(p \ln p + (1 - p) \ln(1 - p))}_{\geq -\ln 2} + \underbrace{(1 - p) \ln \frac{1}{1 - q}}_{\geq 0} \\ &\geq p \ln \frac{1}{q} - \ln 2.\end{aligned}$$

Substituting back, we have

$$\mathbb{E}_{\nu_1}[n^p_{t+1}(I_1)] \leq \left(\frac{25}{\alpha_{\min}} + \ln 2\right)\left(\ln \frac{t}{\mathbb{E}_{\nu_2}[n^p_{t+1}(I_1)]}\right)^{-1} \cdot t \leq \frac{26}{\alpha_{\min}}\left(\ln \frac{t}{\mathbb{E}_{\nu_2}[n^p_{t+1}(I_1)]}\right)^{-1} \cdot t.$$

Specifically, either that $\mathbb{E}_{\nu_2}[n^p_{t+1}(I_1)] \geq \exp(-52/\alpha_{\min})t = \Omega(t)$ or, through direct substitution, $\mathbb{E}_{\nu_1}[n^p_{t+1}(I_1)] \leq t/2$, i.e., $\mathbb{E}_{\nu_1}[n^p_{t+1}(I_2)] = \Omega(t)$. In other words, any algorithm would choose the suboptimal item a linear number of times in at least one of the two problems, and since the items have strictly positive quality gap, this would incur a linear regret. □

E Proof of Thm. 2

Proof. We denote $p_{\min} = \min_{i\in[M],u_t\in\mathcal{U}_{\boldsymbol{s}_t}\in\mathcal{S},t\geq 1} z_i(\mathrm{qb}(u_t,\boldsymbol{s}_t)+\mathrm{rb}+\mathrm{pb}(u_t,\boldsymbol{s}_t))$. Notice that due to boundness of disp_t for all t, $p_{\min}\geq\Omega\big(\frac{1}{M}\big)$.

Next, we let $N_t(I)$ denote the number of times item I was selected by the *ranker* up to time t. That is, $N_t(I)=\sum_{k=1}^{t}\mathbb{1}\{I\in\boldsymbol{s}_k\}$. Note that it is not necessarily true that $N_t(I)=n_t(I)$.

Let $I\in\mathcal{D}, m\in[T]$, and let $\tau_i^*(I)=\min\{t\geq 1: N_t(I)=i\}$. For any $m\in[T]$, we define the coupling $\{(X_i,Y_i)\}_{i=1}^m$, where $\{X_i\}_{i=1}^m$ are iid Bernouli random variables with probability $p_{\min}$ and $Y_i=\mathbb{1}\big\{c_{\tau_i^*(I)}=I\big\}$. Then, for any $a\in\mathbb{R}$,

$$P\left(\sum_{i=1}^{m}Y_i\leq a, N_t(I)=m\right)\leq P\left(\sum_{i=1}^{m}X_i\leq a, N_t(I)=m\right). \tag{7}$$

Denote $m_0=\frac{8b_{\max}}{p_{\min}\alpha_{\min}}\log(\frac{3|\mathcal{D}|}{\delta})$ and let $t\geq m_0$. Then,

$$\begin{aligned}
&P\left(n_t(I)\leq\frac{b_{\max}}{\alpha_{\min}}, N_t(I)\geq m_0\right)\\
&=P\left(\sum_{k=1}^{t}\mathbb{1}\{c_k=I\}\leq\frac{b_{\max}}{\alpha_{\min}}, N_t(I)\geq m_0\right)\\
&\leq\sum_{m=m_0}^{\infty}P\left(\sum_{k=1}^{t}\mathbb{1}\{c_k=I\}\leq\frac{b_{\max}}{\alpha_{\min}}, N_t(I)=m\right) && \text{(Union Bound)}\\
&\leq\sum_{m=m_0}^{\infty}P\left(\sum_{k:I\in S_k}\mathbb{1}\{c_k=I\}\leq\frac{b_{\max}}{\alpha_{\min}}, N_t(I)=m\right)\\
&=\sum_{m=m_0}^{\infty}P\left(\sum_{i=1}^{N_t(I)}\mathbb{1}\big\{c_{\tau_i^*(I)}=I\big\}\leq\frac{b_{\max}}{\alpha_{\min}}, N_t(I)=m\right)\\
&\leq\sum_{m=m_0}^{\infty}P\left(\sum_{i=1}^{N_t(I)}X_i\leq\frac{b_{\max}}{\alpha_{\min}}, N_t(I)=m\right) && \text{(Eq. (7))}\\
&=\sum_{m=m_0}^{\infty}P\left(\sum_{i=1}^{m}X_i\leq\frac{b_{\max}}{\alpha_{\min}}, N_t(I)=m\right)\\
&\leq\sum_{m=m_0}^{\infty}P\left(\sum_{i=1}^{m}X_i\leq\frac{b_{\max}}{\alpha_{\min}}\right)
\end{aligned}$$

By Hoeffdings inequality, since $X_i\overset{\text{iid}}{\sim}\text{Bern}(p_{\min})$,

$$P\left(\sum_{i=1}^{m}X_i-p_{\min}m\leq-\frac{p_{\min}m}{2}\right)\leq\exp\{-m/2\}.$$

Therefore,

$$P\left(\sum_{i=1}^{m}X_i\leq\frac{p_{\min}m}{2}\right)\leq\exp\{-m/2\}.$$

Using the above and the definition of m_0 we get that

$$\begin{aligned}
P\left(n_t(I) \leq \frac{b_{\max}}{\alpha_{\min}}, N_t(I) \geq m_0\right) &\leq \sum_{m=m_0}^{\infty} P\left(\sum_{i=1}^{m} X_i \leq \frac{b_{\max}}{\alpha_{\min}}\right) \\
&\leq \sum_{m=m_0}^{\infty} P\left(\sum_{i=1}^{m} X_i \leq \frac{p_{\min} m}{2}\right) \\
&\leq \sum_{m=m_0}^{\infty} \exp\{-m/2\} \\
&\leq 3\exp\{-m_0/2\} \leq \frac{\delta}{|\mathcal{D}|}.
\end{aligned}$$

By the union bound, for any $t \geq m_0$

$$P\left(\bigcup_{I \in \mathcal{D}} \left\{n_t(I) \leq \frac{b_{\max}}{\alpha_{\min}}, N_t(I) \geq m_0\right\}\right) \leq \delta.$$

By Assumption 1, $\{[\mathrm{pb}_t(u, \boldsymbol{s}_t)]_i < b_i(u, \boldsymbol{s}_t)\} \subseteq \left\{n_t(I_{i,t}) \leq \frac{b_{\max}}{\alpha_{\min}}\right\}$. Therefore, it also holds that

$$P\left(\bigcup_{I \in \mathcal{D}} \{\exists i \leq M : I_{i,t} = I, [\mathrm{pb}_t(u, \boldsymbol{s}_t)]_i < b_i(u, \boldsymbol{s}_t), N_t(I) \geq m_0\}\right) \leq \delta.$$

In other words, w.p. at least $1-\delta$, for all $I \in \mathcal{D}$ and $i \leq M$, if item I is in the slate ($I_{i,t} = I$) and $N_t(I) \geq m_0$, then its popularity bias is saturated ($[\mathrm{pb}_t(u, \boldsymbol{s}_t)]_i = b_i(u, \boldsymbol{s}_t)$). On the other hand, if an item is in the slate and $N_t(I) < m_0$, then $N_t(I)$ increases by 1. Thus, by the pigeonhole principle, the number of rounds such that an item in the slate has $N_t(I) < m_0$ is bounded by $(m_0+1)|\mathcal{D}|$.

Finally, noticing that when all items in the slate are saturated, the quality ranker is optimal, and otherwise, the instantaneous regret is bounded by 1, we get w.p. at least $1-\delta$ that

$$\begin{aligned}
\mathrm{Reg}_{\mathcal{R}_q}(T) &\leq \sum_{t=1}^{T} \mathbb{1}\{\exists i \leq M : [\mathrm{pb}_t(u, \boldsymbol{s}_t)]_i < b_i(u, \boldsymbol{s}_t)\} \\
&\leq \sum_{t=1}^{T} \mathbb{1}\{\exists i \leq M : I_{i,t} = I, N_t(I) < m_0\} \\
&\leq \min\{(m_0+1)|\mathcal{D}|, T\}.
\end{aligned}$$

Plugging in the definition of m_0 and using the fact that $p_{\min} \geq \Omega\left(\frac{1}{M}\right)$ completes the proof.

□

F Proof of Thm. 3

For clarity, we denote

$$v_t(u, \boldsymbol{s}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{i=1}^{M} [\mathrm{qb}_{\boldsymbol{\theta}}(u, \boldsymbol{s})]_i z_i \big(\mathrm{qb}_{\boldsymbol{\theta}}(u, \boldsymbol{s}) + \mathrm{pb}_{t,\boldsymbol{\phi}}(u, \boldsymbol{s}, \boldsymbol{c}_{1:t-1}) + \mathrm{rb}\big).$$

and

$$v_{\mathrm{sat}}(u, \boldsymbol{s}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{i=1}^{M} [\mathrm{qb}_{\boldsymbol{\theta}}(u, \boldsymbol{s})]_i z_i (\mathrm{qb}_{\boldsymbol{\theta}}(u, \boldsymbol{s}) + \boldsymbol{b}_{\boldsymbol{\phi}}(u, \boldsymbol{s}) + \mathrm{rb}).$$

Also, recall that

$$\boldsymbol{s}_t \in \arg\max_{\boldsymbol{s} \in \mathcal{S}} v_{\mathrm{sat}}(u_t, \boldsymbol{s}, \boldsymbol{\theta}_t, \boldsymbol{\phi}_t) + \epsilon_t(\boldsymbol{s}). \tag{8}$$

Let $\mathcal{E}_t^{\mathrm{sat}} = \{\{\mathrm{pb}_t(u, \boldsymbol{s}_t, \boldsymbol{c}_{1:t-1}) = \boldsymbol{b}(u, \boldsymbol{s}_t)\}\}$. Similar to Lemma 1, with probability at least $1 - \delta$, we have that

$$\sum_{t=1}^{T} \mathbb{1}\big\{\big(\mathcal{E}_t^{\mathrm{sat}}\big)^c\big\} \le \frac{8\,|\mathcal{D}|\, M b_{\max}}{\alpha_{\min}} \log\left(\frac{2\,|\mathcal{D}|}{\delta}\right).$$

Then, with probability at least $1 - \delta$,

$$\begin{aligned}
\mathrm{Reg}_{\mathcal{R}_{qp}}(T) &= \sum_{t=1}^{T} \mathbb{E}_{u_t}[(v_t(u_t, \boldsymbol{s}^q, \boldsymbol{\theta}^*, \boldsymbol{\phi}^*) - v_t(u_t, \boldsymbol{s}_t, \boldsymbol{\theta}^*, \boldsymbol{\phi}^*))] \\
&\le \sum_{t=1}^{T} \mathbb{E}_{u_t}\big[(v_t(u_t, \boldsymbol{s}^q, \boldsymbol{\theta}^*, \boldsymbol{\phi}^*) - v_t(u_t, \boldsymbol{s}_t, \boldsymbol{\theta}^*, \boldsymbol{\phi}^*))\mathbb{1}\big\{\mathcal{E}_t^{\mathrm{sat}}\big\}\big] + \sum_{t=1}^{T} 2\mathbb{1}\big\{\big(\mathcal{E}_t^{\mathrm{sat}}\big)^c\big\} \\
&\le \sum_{t=1}^{T} \mathbb{E}_{u_t}[v_{\mathrm{sat}}(u_t, \boldsymbol{s}^q, \boldsymbol{\theta}^*, \boldsymbol{\phi}^*) - v_{\mathrm{sat}}(u_t, \boldsymbol{s}_t, \boldsymbol{\theta}^*, \boldsymbol{\phi}^*)] + \frac{16\,|\mathcal{D}|\, M b_{\max}}{\alpha_{\min}} \log\left(\frac{2\,|\mathcal{D}|}{\delta}\right) \\
&\le \sum_{t=1}^{T} \mathbb{E}_{u_t}[v_{\mathrm{sat}}(u_t, \boldsymbol{s}^q, \boldsymbol{\theta}_t, \boldsymbol{\phi}_t) + \epsilon_t(\boldsymbol{s}^q) - v_{\mathrm{sat}}(u_t, \boldsymbol{s}_t, \boldsymbol{\theta}_t, \boldsymbol{\phi}_t) + \epsilon_t(\boldsymbol{s}_t)] + \frac{16\,|\mathcal{D}|\, M b_{\max}}{\alpha_{\min}} \log\left(\frac{2\,|\mathcal{D}|}{\delta}\right) \\
&\qquad \text{(Lemma 1)} \\
&\le \sum_{t=1}^{T} \mathbb{E}_{u_t}[v_{\mathrm{sat}}(u_t, \boldsymbol{s}_t, \boldsymbol{\theta}_t, \boldsymbol{\phi}_t) + \epsilon_t(\boldsymbol{s}_t) - v_{\mathrm{sat}}(u_t, \boldsymbol{s}_t, \boldsymbol{\theta}_t, \boldsymbol{\phi}_t) + \epsilon_t(\boldsymbol{s}_t)] + \frac{16\,|\mathcal{D}|\, M b_{\max}}{\alpha_{\min}} \log\left(\frac{2\,|\mathcal{D}|}{\delta}\right) \\
&\qquad \text{(Eq. (8))} \\
&= 2\sum_{t=1}^{T} \epsilon_t(\boldsymbol{s}_t) + \frac{16\,|\mathcal{D}|\, M b_{\max}}{\alpha_{\min}} \log\left(\frac{2\,|\mathcal{D}|}{\delta}\right) \\
&\le 2\sum_{t=1}^{T} \left(4\sqrt{M} + \frac{1}{\sqrt{1-\rho_{min}}}\right) \gamma_t(\delta) \sqrt{\mathbb{E}_{t-1}\left[\|\boldsymbol{x}(u, \boldsymbol{s})\|^2_{\boldsymbol{V}_t^{-1}}\right]} + \frac{16\,|\mathcal{D}|\, M b_{\max}}{\alpha_{\min}} \log\left(\frac{2\,|\mathcal{D}|}{\delta}\right) \\
&\le 2\left(4\sqrt{M} + \frac{1}{\sqrt{1-\rho_{min}}}\right) \gamma_T(\delta) \sqrt{T\mathbb{E}_{t-1}\left[\sum_{t=1}^{T} \|\boldsymbol{x}(u, \boldsymbol{s})\|^2_{\boldsymbol{V}_t^{-1}}\right]} + \frac{16\,|\mathcal{D}|\, M b_{\max}}{\alpha_{\min}} \log\left(\frac{2\,|\mathcal{D}|}{\delta}\right) \\
&\qquad \text{(Cauchy–Schwarz ineq.)} \\
&\le \mathcal{O}\left(\left(\sqrt{M} + \frac{1}{\sqrt{1-\rho_{min}}}\right) \gamma_T(\delta) \sqrt{(d_q + d_p) T \log\left(1 + \frac{T}{\lambda(d_q + d_p)}\right)}\right)
\end{aligned}$$

where the last inequality is due to Abbasi-Yadkori et al. [2011] (Lemma 11). Finally letting $\lambda \leq \mathcal{O}\left(\frac{1}{\sqrt{L}}\right)$ yields the desired result.

$$\begin{aligned}\mathrm{Reg}_{\mathcal{R}_{qp}}(T) &\leq \mathcal{O}\left(M^2\left(\sqrt{M}+\frac{1}{\sqrt{1-\rho_{min}}}\right)\sqrt{\log(1/\delta)+Md\log\left(1+\frac{TL}{d}\right)}\sqrt{dT\log\left(1+\frac{TL}{d}\right)}\right)\\ &\leq \mathcal{O}\left(M^{2.5}d\sqrt{T}\left(\sqrt{M}+\frac{1}{\sqrt{1-\rho_{min}}}\right)\sqrt{\log(1/\delta)}\log\left(1+\frac{TL}{d}\right)\right)\end{aligned}$$

Lemma 1. *For any $\boldsymbol{s}\in\mathcal{S}, t\in[T], \delta\in(0,1)$, with probability at least $1-\delta$*

$$|\mathbb{E}_u[v_{sat}(u,\boldsymbol{s},\boldsymbol{\theta}^*,\boldsymbol{\phi}^*)-v_{sat}(u,\boldsymbol{s},\boldsymbol{\theta}_t,\boldsymbol{\phi}_t)]|\leq\epsilon_t(\boldsymbol{s})$$

Proof. We have that

$$\begin{aligned}&|\mathbb{E}_u[v_{\text{sat}}(u,\boldsymbol{s},\boldsymbol{\theta}^*,\boldsymbol{\phi}^*)-v_{\text{sat}}(u,\boldsymbol{s},\boldsymbol{\theta}_t,\boldsymbol{\phi}_t)]|\\ &=\left|\mathbb{E}_u\left[\sum_{i=1}^M[\mathrm{qb}_{\boldsymbol{\theta}^*}(u,\boldsymbol{s})]_i z_i(\mathrm{qb}_{\boldsymbol{\theta}^*}(u,\boldsymbol{s})+\boldsymbol{b}_{\boldsymbol{\phi}^*}(u,\boldsymbol{s})+\mathrm{rb})-\left[\mathrm{qb}_{\boldsymbol{\theta}_t}(u,\boldsymbol{s})\right]_i z_i\left(\mathrm{qb}_{\boldsymbol{\theta}_t}(u,\boldsymbol{s})+\boldsymbol{b}_{\boldsymbol{\phi}_t}(u,\boldsymbol{s})+\mathrm{rb}\right)\right]\right|\\ &\leq\left|\mathbb{E}_u\left[\sum_{i=1}^M[\mathrm{qb}_{\boldsymbol{\theta}^*}(u,\boldsymbol{s})]_i\left(z_i(\mathrm{qb}_{\boldsymbol{\theta}^*}(u,\boldsymbol{s})+\boldsymbol{b}_{\boldsymbol{\phi}^*}(u,\boldsymbol{s})+\mathrm{rb})-z_i\left(\mathrm{qb}_{\boldsymbol{\theta}_t}(u,\boldsymbol{s})+\boldsymbol{b}_{\boldsymbol{\phi}_t}(u,\boldsymbol{s})+\mathrm{rb}\right)\right)\right]\right|\\ &\quad+\left|\mathbb{E}_u\left[\sum_{i=1}^M\left([\mathrm{qb}_{\boldsymbol{\theta}^*}(u,\boldsymbol{s})]_i-\left[\mathrm{qb}_{\boldsymbol{\theta}_t}(u,\boldsymbol{s})\right]_i\right)z_i\left(\mathrm{qb}_{\boldsymbol{\theta}_t}(u,\boldsymbol{s})+\boldsymbol{b}_{\boldsymbol{\phi}_t}(u,\boldsymbol{s})+\mathrm{rb}\right)\right]\right|\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(Triangle Ineq.)}\end{aligned}$$

By Lemma 3,

$$\begin{aligned}&\left|\mathbb{E}_u\left[\sum_{i=1}^M[\mathrm{qb}_{\boldsymbol{\theta}^*}(u,\boldsymbol{s})]_i\left(z_i(\mathrm{qb}_{\boldsymbol{\theta}^*}(u,\boldsymbol{s})+\boldsymbol{b}_{\boldsymbol{\phi}^*}(u,\boldsymbol{s})+\mathrm{rb})-z_i\left(\mathrm{qb}_{\boldsymbol{\theta}_t}(u,\boldsymbol{s})+\boldsymbol{b}_{\boldsymbol{\phi}_t}(u,\boldsymbol{s})+\mathrm{rb}\right)\right)\right]\right|\\ &\leq 4\sqrt{M}\gamma_t(\delta)\mathbb{E}_u\left[\|\boldsymbol{x}(u,\boldsymbol{s})\|_{\boldsymbol{V}_t^{-1}}\right]\\ &\leq 4\sqrt{M}\gamma_t(\delta)\sqrt{\mathbb{E}_u\left[\|\boldsymbol{x}(u,\boldsymbol{s})\|^2_{\boldsymbol{V}_t^{-1}}\right]}.\end{aligned}$$

By Hölder's inequality and Lemma 2,

$$\begin{aligned}&\left|\mathbb{E}_u\left[\sum_{i=1}^M\left([\mathrm{qb}_{\boldsymbol{\theta}^*}(u,\boldsymbol{s})]_i-\left[\mathrm{qb}_{\boldsymbol{\theta}_t}(u,\boldsymbol{s})\right]_i\right)z_i\left(\mathrm{qb}_{\boldsymbol{\theta}_t}(u,\boldsymbol{s})+\boldsymbol{b}_{\boldsymbol{\phi}_t}(u,\boldsymbol{s})+\mathrm{rb}\right)\right]\right|\\ &\leq\left\|\mathbb{E}_u\left[\mathrm{qb}_{\boldsymbol{\theta}^*}(u,\boldsymbol{s})-\mathrm{qb}_{\boldsymbol{\theta}_t}(u,\boldsymbol{s})\right]\right\|_\infty\left\|\mathbb{E}_u\left[\boldsymbol{z}\left(\mathrm{qb}_{\boldsymbol{\theta}_t}(u,\boldsymbol{s})+\boldsymbol{b}_{\boldsymbol{\phi}_t}(u,\boldsymbol{s})+\mathrm{rb}\right)\right]\right\|_1\\ &=\left\|\mathbb{E}_u\left[\mathrm{qb}_{\boldsymbol{\theta}^*}(u,\boldsymbol{s})-\mathrm{qb}_{\boldsymbol{\theta}_t}(u,\boldsymbol{s})\right]\right\|_\infty\\ &\leq\frac{\gamma_t(\delta)}{\sqrt{1-\rho_{min}}}\sqrt{\mathbb{E}_u\left[\|\boldsymbol{x}(u,\boldsymbol{s})\|^2_{\boldsymbol{V}_t^{-1}}\right]}.\end{aligned}$$

Combining the above we get the desired result. □

Lemma 2. *For $t>1$*

$$\left|\mathbb{E}_{t-1}\left[[\mathrm{qb}_{\boldsymbol{\theta}^*}(u,\boldsymbol{s})]_i-\left[\mathrm{qb}_{\boldsymbol{\theta}_t}(u,\boldsymbol{s})\right]_i\right]\right|\leq\frac{\gamma_t(\delta)}{\sqrt{1-\rho_{min}}}\sqrt{\mathbb{E}_{t-1}\left[\|\boldsymbol{x}(u,\boldsymbol{s})\|^2_{\boldsymbol{V}_t^{-1}}\right]}$$

Proof. For any $i\in[M], \boldsymbol{s}\in\mathcal{S}, t\in[T]$,

$$\begin{aligned}\left|\mathbb{E}_{t-1}\left[[\mathrm{qb}_{\boldsymbol{\theta}^*}(u,\boldsymbol{s})]_i-\left[\mathrm{qb}_{\boldsymbol{\theta}_t}(u,\boldsymbol{s})\right]_i\right]\right|&=\left|\mathbb{E}_{t-1}\left[\boldsymbol{x}_q^T(u,\boldsymbol{s})\boldsymbol{\theta}_i^*-\boldsymbol{x}_q^T(u,\boldsymbol{s})\boldsymbol{\theta}_{i,t}\right]\right|\\ &=\left|\mathbb{E}_{t-1}\left[\boldsymbol{x}^T(u,\boldsymbol{s})\boldsymbol{D}_{d_q,d_p}(\boldsymbol{\psi}_i^*-\boldsymbol{\psi}_{i,t})\right]\right|\\ &\leq\mathbb{E}_{t-1}\left[\left|\boldsymbol{x}^T(u,\boldsymbol{s})\boldsymbol{D}_{d_q,d_p}(\boldsymbol{\psi}_i^*-\boldsymbol{\psi}_{i,t})\right|\right]\\ &\leq\mathbb{E}_{t-1}\left[\|\boldsymbol{x}(u,\boldsymbol{s})\|_{\boldsymbol{V}_t^{-1}}\left\|\boldsymbol{D}_{d_q,d_p}(\boldsymbol{\psi}_i^*-\boldsymbol{\psi}_{i,t})\right\|_{\boldsymbol{V}_t}\right]\end{aligned}$$

By Assumption 3, we have that

$$\begin{aligned}
\mathbb{E}_{t-1}\left[\boldsymbol{D}_{d_q,d_p}\boldsymbol{V}_t\boldsymbol{D}_{d_q,d_p}\right] &= \boldsymbol{D}_{d_q,d_p}\left(\lambda\boldsymbol{I}_{d_p+d_q} + \mathbb{E}_{t-1}\left[\sum_{k=1}^{t-1}\boldsymbol{x}(u_k,\boldsymbol{s}_k)\boldsymbol{x}^T(u_k,\boldsymbol{s}_k)\right]\right)\boldsymbol{D}_{d_q,d_p} \\
&= \lambda\boldsymbol{D}_{d_q,d_p} + \mathbb{E}_{t-1}\left[\begin{pmatrix}\boldsymbol{\Sigma_{qq}}(u,\boldsymbol{s}) & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\right] \\
&\preceq \lambda\boldsymbol{I}_{d_q+d_p} + \frac{1}{1-\rho}\mathbb{E}_{t-1}\left[\begin{pmatrix}\boldsymbol{\Sigma_{qq}}(u,\boldsymbol{s}) & \boldsymbol{\Sigma_{qp}}(u,\boldsymbol{s})\\ \boldsymbol{\Sigma_{pq}}(u,\boldsymbol{s}) & \boldsymbol{\Sigma_{pp}}(u,\boldsymbol{s})\end{pmatrix}\right] \\
&\preceq \frac{1}{1-\rho}\boldsymbol{V}_t. \qquad (9)
\end{aligned}$$

Then,

$$\begin{aligned}
&\mathbb{E}_{t-1}\left[\|\boldsymbol{x}(u,\boldsymbol{s})\|_{\boldsymbol{V}_t^{-1}}\left\|\boldsymbol{D}_{d_q,d_p}(\boldsymbol{\psi}_i^* - \boldsymbol{\psi}_{i,t})\right\|_{\boldsymbol{V}_t}\right] \\
&\leq \sqrt{\mathbb{E}_{t-1}\left[\|\boldsymbol{x}(u,\boldsymbol{s})\|^2_{\boldsymbol{V}_t^{-1}}\right]\mathbb{E}_{t-1}\left[\left\|\boldsymbol{D}_{d_q,d_p}(\boldsymbol{\psi}_i^* - \boldsymbol{\psi}_{i,t})\right\|^2_{\boldsymbol{V}_t}\right]} \qquad \text{(Cauchy–Schwarz ineq.)} \\
&\leq \frac{1}{\sqrt{1-\rho_{min}}}\sqrt{\mathbb{E}_{t-1}\left[\|\boldsymbol{x}(u,\boldsymbol{s})\|^2_{\boldsymbol{V}_t^{-1}}\right]\mathbb{E}_{t-1}\left[\left(\|\boldsymbol{\psi}_i^* - \boldsymbol{\psi}_{i,t}\|^2_{\boldsymbol{V}_t}\right)\right]} \qquad \text{(Eq. (9))} \\
&\leq \frac{1}{\sqrt{1-\rho_{min}}}\sqrt{\mathbb{E}_{t-1}\left[\|\boldsymbol{x}(u,\boldsymbol{s})\|^2_{\boldsymbol{V}_t^{-1}}\right]\mathbb{E}_{t-1}\left[\|\boldsymbol{\psi}_i^* - \boldsymbol{\psi}_t\|^2_{\boldsymbol{I}_M\otimes\boldsymbol{V}_t}\right]} \\
&\leq \frac{\gamma_t(\delta)}{\sqrt{1-\rho_{min}}}\sqrt{\mathbb{E}_{t-1}\left[\|\boldsymbol{x}(u,\boldsymbol{s})\|^2_{\boldsymbol{V}_t^{-1}}\right]}. \qquad \text{(Lemma 4)}
\end{aligned}$$

□

This completes the proof.

G Useful Lemmas

Lemma 3. *With probability at least* $1-\delta$

$$\left|\mathbb{E}_u\left[\sum_{i=1}^{M}[\mathrm{qb}_{\boldsymbol{\theta}^*}(u,\boldsymbol{s})]_i\left(z_i(\mathrm{qb}_{\boldsymbol{\theta}^*}(u,\boldsymbol{s})+\boldsymbol{b}_{\boldsymbol{\phi}^*}(u,\boldsymbol{s})+\mathrm{rb})-z_i\big(\mathrm{qb}_{\boldsymbol{\theta}_t}(u,\boldsymbol{s})+\boldsymbol{b}_{\boldsymbol{\phi}_t}(u,\boldsymbol{s})+\mathrm{rb}\big)\right)\right]\right| \leq \sqrt{M}\gamma_t(\delta)\mathbb{E}_u\left[\|\boldsymbol{x}(u,\boldsymbol{s})\|_{\boldsymbol{V}_t^{-1}}\right].$$

Proof. The proof follows the same step as the proof in Appendix C.2 of Amani and Thrampoulidis [2021], using the fact that $\|\mathrm{qb}\|_2 = \sqrt{\sum_{i=1}^{M}\|\mathrm{qb}_i\|_2^2} = \sqrt{M}$, and, e.g., Eq. (29) of Amani and Thrampoulidis [2021] for which $\sup\lambda_{\max}(\boldsymbol{A})\cdot\inf 1/\lambda_{\min}(\boldsymbol{A}) \leq e^2(1+Me)^2 \leq 2e^4M^2$. Note that this paper uses a different definition of γ_t than in Amani and Thrampoulidis [2021]. □

Lemma 4. *With probability at least* $1-\delta$

$$\mathbb{E}_u\left[\|\boldsymbol{\psi}^*-\boldsymbol{\psi}_t\|^2_{\boldsymbol{I}_M\otimes\boldsymbol{V}_t}\right] \leq \gamma_t(\delta)$$

Proof. Follows from Lemma 14 of Amani and Thrampoulidis [2021]. □