Preselection Bandits

Viktor Bengs[1] **Eyke Hüllermeier**[1]

Abstract

In this paper, we introduce the Preselection Bandit problem, in which the learner preselects a subset of arms (choice alternatives) for a user, which then chooses the final arm from this subset. The learner is not aware of the user's preferences, but can learn them from observed choices. In our concrete setting, we allow these choices to be stochastic and model the user's actions by means of the Plackett-Luce model. The learner's main task is to preselect subsets that eventually lead to highly preferred choices. To formalize this goal, we introduce a reasonable notion of regret and derive lower bounds on the expected regret. Moreover, we propose algorithms for which the upper bound on expected regret matches the lower bound up to a logarithmic term of the time horizon.

1. Introduction

The setting of preference-based multi-armed bandits or *dueling bandits* (Busa-Fekete et al., 2018) is a generalization of the standard stochastic multi-armed bandit (MAB) problem (Lattimore & Szepesvári, 2020). Instead of assuming numerical rewards of individual arms (choice alternatives), the former is based on pairwise preferences between arms. In this paper, we introduce the Preselection Bandit (or simply Pre-Bandit) problem, which is closely related to the preference-based setting, especially to the recent variant of *battling bandits* (Saha & Gopalan, 2018).

Our setting involves an *agent* (learner), which preselects a subset of arms, and a *selector* (a human user or another algorithm), which then chooses the final arm from this subset. This setting is motivated by various practical applications. In information retrieval, for example, the role of the agent is played by a search engine, and the selector is the user who seeks a certain information. Another example is online advertising, where advertisements recommended to users can be seen as a preselection. As a concrete application, we are currently working on the problem of algorithm (pre-) selection (Kerschke et al., 2019), where the (presumably) best-performing algorithm needs to be chosen from a pool of candidates.

In the beginning, the agent is not aware of the selector's preferences. However, the choices made by the latter reveal information about these preferences, from which the agent can learn. Due to time constraints, information asymmetry, or other reasons, we do not assume the selector to act perfectly, which means that it may miss the actually best among the preselected arms. In web search, for example, a user clicks on links based on limited information such as snippets, but without knowing the full content behind. Likewise, in algorithm selection, the final choice might be made on the basis of a cross-validation study, i.e., estimated performances that not guarantee the identification of the truly best algorithm. By modeling the selector's actions by means of the Plackett-Luce (PL) model (Luce, 1959; Plackett, 1975), we allow some randomness in the process of decision making. The agent's main task is to preselect subsets that eventually lead to highly preferred choices. To formalize this goal, we introduce a reasonable notion of regret based on utilities of preselected subsets, where an arm's (latent) utility is weighted with the probability of choosing this arm from the subset. In particular, this allows for capturing decision-making biases of users as studied intensively in behavioral economics or psychology.

We study two variants of the problem. In the first variant, which we call *restricted* Pre-Bandit problem, the size of the preselection is predefined and fixed throughout. In the second variant, the *flexible* Pre-Bandit problem, the agent is allowed to adjust the size of the preselection in every round. For these settings, we derive lower bounds on the expected regret. Moreover, for both scenarios, we propose active learning algorithms for which the upper bound on expected regret matches the lower bound (possibly) up to a logarithmic term of the time horizon.

We discuss related work in Section 2. In Section 3, we introduce the notation used throughout the paper, and also give a concise review of the PL model and some of its properties. In Section 4, the Pre-Bandit problem is formally introduced, together with a reasonable notion of regret, for which lower

[1] Heinz Nixdorf Institute and Department of Computer Science, Paderborn University, Germany. Correspondence to: Viktor Bengs <viktor.bengs@upb.de>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

bounds with respect to the time horizon are verified. Near-optimal algorithms for the two variants of the Pre-Bandit problem are provided in Section 5. We devote Section 6 to a simulation study demonstrating the usefulness and efficiency of our algorithms. Finally, Section 7 summarizes our results and discusses directions for future work. All proofs of the theoretical results are deferred to the supplements.

2. Related Work

Bandit problems with the possibility of using more than one arm at a time have been considered in various ways in the literature. However, as will be detailed in the following, none of the previous works encompasses the problem considered in this paper. Due to the specific meaning of the subset choice as a preselection in practical applications, this also justifies a new name for the setting.

A variant of the MAB problem is the combinatorial bandit problem (Cesa-Bianchi & Lugosi, 2012; Kveton et al., 2015), in which the learner chooses a subset of arms in each time step, and then observes quantitative feedback, either in the form of rewards of each single arm (semi-bandit feedback) or the total sum of the rewards (bandit feedback) for the arms in the chosen subset. This differs fundamentally from our setting, in which no quantitative feedback is ever observed; instead, only qualitative feedback is provided, i.e., which arm is picked in a subset.

Qualitative forms of feedback for multiple arm choices at a time is considered in the realm of dueling, multi-dueling (Sui et al., 2017; Busa-Fekete et al., 2018) or battling bandits (Saha & Gopalan, 2018). The flexible Pre-Bandit problem has obvious connections to the latter settings, with the freedom of adjusting the size of comparison for each time instance and can be interpreted as a combinatorial bandit problem with qualitative feedback. Saha & Gopalan (2019a) investigate the effect of this flexibility in an active PAC-framework for finding the best arm under the PL model, while the active top-k-arm identification problem in this model is studied by Chen et al. (2018a). Recently, this scenario was considered in terms of a regret minimization problem with top-m-ranking feedback by Saha & Gopalan (2019b) for a straightforward extension of the dueling bandit notion of regret and not regarding the “value” of a subset in its entirety as we do (cf. Section 4.1).

Moreover, the algorithms suggested in Saha & Gopalan (2018) are not applicable within our scenario for the restricted Pre-Bandit problem, as either the algorithms are focusing on the linear-subset choice model, which is fundamentally different from our choice model (cf. Section 3) or the algorithms allow replicates of the same arm within a chosen subset, which is inadmissible within our scenario (and in many practical applications as well).

The Pre-Bandit problem also reveals parallels to the *Dynamic Assortment Selection* (DAS) problem (Caro & Gallien, 2007), where a retailer seeks to find an optimal subset of his/her available items (or products) in an online manner, so as to maximize the expected revenue (or equivalently minimize the expected regret). The DAS problem under the multinomial logit model, which is also known as the *MNL-Bandits* problem (Rusmevichientong et al., 2010; Sauré & Zeevi, 2013; Agrawal et al., 2016; 2017; Wang et al., 2018; Chen et al., 2018b) is especially close to our framework, as the corresponding concept of regret shares similarities with our definition of regret.

However, our problem can rather be seen as complementary, since we do not assume a priori known revenues for each item or any revenues at all. While this might be natural for the retail management problem, it is arguably less so for applications we have in mind, such as recommendation systems or algorithm (pre-)selection. In addition, we introduce a parameter in our setting that allows the learner to adjust its preselections with respect to the preciseness of the user’s selections. This might also be an interesting direction for future work for the DAS problem. Finally, to demonstrate the inappropriateness of the DAS algorithms for the restricted Pre-Bandit problem, we employ some of the algorithms in our experimental study.

Another quite related branch of research is the so-called *stochastic click model* (Zoghi et al., 2017; Lattimore et al., 2018), where a list of l items is presented to the selector in each iteration. Scanning the list from the top to the bottom, there is a certain probability that the selector chooses the item at the current position, or otherwise continues searching (eventually perhaps not choosing any item). Thus, in contrast to our setting, the explicit order of the arms within a subset (or list) is a relevant aspect. Further, the resulting learning task boils down to finding the l most attractive items, as these provably constitute the optimal list in this scenario (which is not necessarily the case for our setting).

3. Preliminaries

3.1. Basic Setting and Notation

We formalize our problem in the setting of preference-based multi-armed bandits (Busa-Fekete et al., 2018), which proceeds from a set of n arms, each of which is considered as a choice alternative (item, option). We identify the arms by the index set $[n] := \{1, \ldots, n\}$, where $n \in \mathbb{N}$ is arbitrary but fixed. Moreover, we assume a total preference order $\succ$, where $i \succ j$ means that the ith is preferred to the jth arm.

Let $\mathbb{A}_l$ be the set of all l-sized subsets of $[n]$ and $\mathbb{A}_{full} := \cup_{l=1}^{n} \mathbb{A}_l$. Moreover, let $\mathbb{S}_n$ be the symmetric group on $[n]$, the elements of which we refer to as *rankings*: each $\mathbf{r} \in \mathbb{S}_n$ defines a ranking in the form of a total order of the arms

$[n]$, with $\mathbf{r}(i)$ the position of arm i. We assume that $\mathbb{S}_n$ is equipped with a probability distribution $\mathbb{P} : \mathbb{S}_n \to [0,1]$. For an integer $l > 1$ and a set of arms $\{i_1, i_2, \ldots, i_l\} \subseteq [n]$, the probability that i_1 is the most preferred among this set is given by

$$q_{i_1,\ldots,i_l} := \sum\nolimits_{\mathbf{r}\in\mathbb{S}_n:\mathbf{r}(i_1)=\min(\mathbf{r}(i_1),\ldots,\mathbf{r}(i_l))} \mathbb{P}(\mathbf{r}) \ . \qquad (1)$$

3.2. The Plackett-Luce Model

The Plackett-Luce (PL) model (Plackett, 1975; Luce, 1959) is a parametric distribution on the symmetric group $\mathbb{S}_n$ with parameter $\theta = (\theta_1, \ldots, \theta_n)^T \in \mathbb{R}_+^n$, where each component θ_k corresponds to the strength of an arm k, which we will refer to as *score parameter.* The probability of a ranking $\mathbf{r} \in \mathbb{S}_n$ under the PL model is

$$\mathbb{P}_\theta(\mathbf{r}) = \prod\nolimits_{i=1}^{n} \frac{\theta_{\mathbf{r}^{-1}(i)}}{\theta_{\mathbf{r}^{-1}(i)} + \ldots + \theta_{\mathbf{r}^{-1}(n)}} \ , \qquad (2)$$

where $\mathbf{r}^{-1}(i)$ denotes the index of the arm on position i. According to (2), PL models a stage-wise construction of a ranking, where in each round, the item to be put on the next position is chosen with a probability proportional to its strength. As a model of discrete choice, the PL distribution has a strong theoretical motivation. For example, it is the only model that satisfies the Luce axiom of choice (Luce, 1959), including independence from irrelevant alternatives (*ILA property*, see (Alvo & Yu, 2014)). Besides, it has a number of appealing mathematical properties. For instance, there is a simple expression for the l-wise marginals in (1):

$$q_{i_1,\ldots,i_l} = \frac{\theta_{i_1}}{\theta_{i_1} + \theta_{i_2} + \ldots + \theta_{i_l}} \qquad (3)$$

This probability is identical to the popular Multinomial Logit (MNL) model, which is a discrete choice probability model considered in various frameworks (Train, 2009). For our purposes, the use of the relative scores

$$O_{i,j} := \frac{\theta_i}{\theta_j}, \qquad i,j \in [n], \qquad (4)$$

will turn out to be advantageous, as they are directly affected by the ILA property of the PL model. Indeed, for $i,j \in [n]$, let $S_{i,j} \in \mathbb{A}_l$ be such that $i,j \in S_{i,j}$. Furthermore, define $S_{-i,j} := S_{i,j} \backslash \{i\}$ and similarly $S_{i,-j} := S_{i,j} \backslash \{j\}$ for $i,j \in [n]$. Then, for any such a set $S_{i,j}$, (3) and (4) imply

$$O_{i,j} = \frac{\theta_i}{\theta_j} = \frac{\theta_i}{\theta_j} \cdot \frac{\sum_{t\in S_{i,j}} \theta_t}{\sum_{t\in S_{i,j}} \theta_t} = \frac{q_{i,S_{-i,j}}}{q_{j,S_{i,-j}}} \ .$$

Without restricting the parameter space $\Theta = \{\theta \in \mathbb{R}_+^n\}$, the PL model in (2) is not (statistically) identifiable, as $\theta \in \mathbb{R}_+^n$ and $\tilde{\theta} = C\,\theta$ for any constant $C > 0$ lead to the same models, i.e. $\mathbb{P}_\theta = \mathbb{P}_{\tilde{\theta}}$. Restricting the parameter space by assuming some normalization condition on the score parameters fixes this issue. Thus, we consider as parameter space the (restricted) unit square w.r.t. the infinity norm,

$$\Theta = \Big\{ \theta = (\theta_1, \ldots, \theta_n)^T \in [\theta_{min}, 1]^n \mid \theta_{min} \in (0,1), \\ \theta_{max} := \max_i \theta_i = 1 \Big\},$$

which leads to an identifiable statistical model $(\mathbb{P}_\theta)_{\theta\in\Theta}$ and naturally yields a normalization of each individual score parameter easing the fast grading of an arm's utility.

For technical reasons, we additionally exclude models that allow scores below a certain threshold θ_{min} (which will be a small constant), as the relative scores in (4) are then well-defined for any pair $(i,j) \in [n]^2$.

3.3. Degree of Preciseness

In our setting, we model the score parameter θ as

$$\theta_i = v_i^\gamma, \quad \forall i \in [n], \qquad (5)$$

where $v_i \in \mathbb{R}_+$ represents the (latent) utility of arm i, while $\gamma \in (0, \infty)$ represents the degree of preciseness of the user's selections: The higher γ, the more (2) resembles a point-mass distribution on the ranking modeling a precise selector that is always able to identify the best arm, while the lower γ, the more (2) resembles a uniform distribution modeling a selector acting purely at random. The effect of γ on the l-marginals in (3) is quite similar.

Note that $v_1, \ldots, v_n, \gamma$ are not separately identifiable (c.f. Section 3.2 in (Train, 2009)), but $\theta_1, \ldots, \theta_n$ are identifiable under our assumptions on the parameter space Θ above. Hence, by fixing γ, the latent utilities $v_1, \ldots, v_n$ are guaranteed to be identifiable.

4. The Pre-Bandit Problem

The considered online learning problem proceeds over a finite time horizon T. For each time instance $t \in [T]$, the agent (i.e., the learner) suggests a subset $S_t \in \mathbb{A}$, where $\mathbb{A}$ is the action space. The agent's action S_t is based on its observations so far. As a new piece of information, it observes the selector's choice (i.e., the user or the environment) of an arm i_t among the offered subset S_t (with probability $q_{i_t, S_t\backslash\{i_t\}}$ given by (3)).

Suppose $r : \mathbb{A} \to \mathbb{R}_+$ is a suitable regret function (to be defined in the next section below). The goal of the learner resp. the agent is to preselect the available arms by means of subsets S_t in every time instance t such that the *expected cumulative regret* over the time horizon, that is $\mathbb{E}_\theta \sum_{t=1}^{T} r(S_t)$ with $\theta \in \Theta$, is minimized. The problem is analyzed for two possible characteristics of the action space:

- (Restricted Preselection) $\mathbb{A} = \mathbb{A}_l$, i.e., a preselection consists of exactly l many arms, where l is a fixed integer strictly greater than one.
- (Flexible Preselection) $\mathbb{A} = \mathbb{A}_{full}$, i.e., a preselection can be any non-empty subset of $[n]$.

In the following, we introduce sensible notions of regret for the considered problem setting. The key question we then address is the following: What is a good preselection to present the selector? Moreover, we provide a lower bound on the related expected cumulative regret.

4.1. Regret Definition

Assuming the selector to behave according to the PL model with score parameter (5), the expected utility of suggesting S is given by

$$\begin{aligned}\mathrm{U}(S) := \mathrm{U}(S; v, \gamma) &= \quad_{i \in S} v_i \cdot q_{i, S \setminus \{i\}} \\ &= \frac{\quad_{i \in S} v_i^{1+\gamma}}{\quad_{i \in S} v_i^{\gamma}}.\end{aligned} \tag{6}$$

Indeed, if $i_t \in [n]$ is the chosen arm at time t, then

$$\mathbb{E}(v_{i_t} \,|\, S) = \quad_{i \in S} v_i \cdot \mathbb{P}(i \text{ is chosen} \,|\, S) = \mathrm{U}(S).$$

Hence, the corresponding optimal preselection is

$$S^* \in \begin{cases} \arg\max_{S \subseteq [n],\, |S| = l} \mathrm{U}(S), & \text{if } \mathbb{A} = \mathbb{A}_l, \\ \arg\max_i \ \theta_i, & \text{if } \mathbb{A} = \mathbb{A}_{full}. \end{cases} \tag{7}$$

The (instantaneous) regret suffered by the selector is anticipated by the agent through

$$r(S) := \mathrm{U}(S^*) - \mathrm{U}(S), \quad S \in \mathbb{A}. \tag{8}$$

Thus, if $S_1, \ldots, S_T$ are suggested for times 1 to T, respectively, the corresponding cumulative regret over T is

$$\mathcal{R}(T) := \quad_{t=1}^{T} r(S_t) = \quad_{t=1}^{T} \big(\mathrm{U}(S^*) - \mathrm{U}(S_t)\big). \tag{9}$$

Remark 1 (Relations to dueling bandits and battling bandits). Note that the optimal subset for $\mathbb{A} = \mathbb{A}_{full}$, i.e., for the flexible Pre-Bandit problem, always consists of the items whose score parameters equal the overall highest score θ_{max}. Thus, like for the dueling bandits and battling bandits problem, the goal is to find the best arm(s). However, whilst in the latter settings only pairwise resp. fixed l-wise comparisons of arms are observed, we allow to draw comparisons of arbitrary size. In addition, the restricted Pre-Bandit problem can be interpreted as a dueling resp. battling bandit problem. Compared to the latter, however, the notion of regret has a more natural meaning in our setting. This is due to the different semantics of a selection of a pair (or any subset) of arms, which is a preselection that eventually leads to a concrete choice. In other words, the regret in (8) focuses on the perceived preference of a subset by regarding the "value" of a subset in its entirety.

Remark 2 (No-choice option). In the related branches of literature (cf. Section 2), it is common to assume an additional choice alternative which represents the possibility of the user choosing none of the alternatives in the preselection. Formally, this can be expressed in our setting by extending the n-dimensional parameter space Θ to $n+1$ dimensions by augmenting it with a dummy score parameter $\theta_0 \in (\theta_{min}, 1]$ representative for this no-choice option. As this option is always available, this dummy item is part of every preselection, and consequently its latent utility affects only the choice probabilities in (6). However, although we refrain from incorporating the no-choice alternative in this paper, it is straightforward to show similar lower bounds as below for this problem and to modify the suggested algorithms to this scenario.

4.2. Most preferred Subsets

One tempting question is how the most preferred subsets look like, given our definition of regret. As already mentioned, the optimal preselection S^* for the flexible Pre-Bandit variant consists of the items with the same highest score parameter. However, in the restricted Pre-Bandit variant, the optimal preselection does not necessarily consist of the l items with the highest scores in general, as the following examples demonstrate.

Example 1. *In Table 1, we provide three problem instances with fixed degree of preciseness $\gamma = 1$ for $n = 5$ and the corresponding expected scores of (the relevant) 3-sized subsets of $[n]$. In the first instance, where one arm has a much higher utility than the remaining ones, it is favorable to suggest this high utility arm together with the arms having smallest utility. This is due to the large differences between the utilities, so that the selector will take the best arm with a sufficiently high probability. Roughly speaking, the best strategy for the agent is to make the problem for the selector as easy as possible. The second instance is different, as the optimal preselection for the agent now consists of the*

Table 1. Problem instances with different optimal subsets (indicated in bold font) for the regret in (8) with $n = 5$ and $l = 3$ (omitted subsets had smaller utilities throughout).

S	$\{1,2,3\}$	$\{1,2,5\}$	$\{1,3,5\}$	$\{1,4,5\}$	$\{2,3,4\}$
	$v = (1\,, 0.122\,, 0.044\,, 0.037\,, 0.017), \gamma = 1$				
$\mathrm{U}(S)$	0.872	0.891	0.945	**0.951**	0.0896
	$v = (1\,, 0.681\,, 0.572\,, 0.543\,, 0.399), \gamma = 1$				
$\mathrm{U}(S)$	**0.795**	0.780	0.754	0.749	0.604
	$v = (1\,, 0.681\,, 0.572\,, 0.543\,, 0.171), \gamma = 1$				
$\mathrm{U}(S)$	0.795	**0.806**	0.778	0.773	0.604

top-3 arms with the highest scores. This comes with a non-negligible probability of missing the optimal arm, however, since the runner-up arms are sufficiently strong, the regret can be tolerated. On the other hand, adding a poor arm would be suboptimal, as one cannot be certain enough that it will not be taken. But by reducing the score for the worst arm notably as in the third instance, the worst arm substitutes the third best, as then the best item can again be better distinguished from the suboptimal ones inside the optimal subset.

As suggested by this example, a reasonable strategy is to compose the preselection of subsets of best and worst arms, respectively. In fact, we show in the supplementary material (Section D) that the optimal subsets for the restricted Pre-Bandit problem are always composed of best and worst arms with the overall best arm(s) mandatory inside the optimal subset.

The obvious rationale of adding a strong arm is to guarantee a reasonably high utility, whereas a poor arm merely serves as a decoy to increase the probability of choosing the best arm. Such effects are known in the literature on decision theory as the *attraction effect* or the *decoy effect*, see (Dimara et al., 2017) and references therein. In particular, our definition of regret is able to capture this effect and consequently emphasizes that our regret aims at penalizing difficult decisions for the selector in the restricted case.

However, the manifestation of the decoy effect and consequently its demand for the application at hand, can be steered by the learner through fixing the degree of preciseness γ as illustrated by the next example.

Example 2. *In Table 2, we investigate the effect of the degree of preciseness γ for the third instance of Example 1. In the first instance, where the degree of preciseness is moderate, i.e., $\gamma = 1$, the attraction effect is still present. But by increasing the degree of preciseness to $\gamma = 20$ as in the second instance, the (rounded) utilities of all subsets containing the best item are the same. For such a problem instance, it is enough to provide the best item in the preselection, as the user's probability of choosing the best item is sufficiently high in these cases and the best arm will not be missed. In the last instance, the utility monotonically decreases with the sum of scores of the subset S if the degree of preciseness is reduced to $\gamma = 0.05$. This is due to a throughout non-negligible probability for choosing the worst item inside the preselected subset S. Thus, in order to dampen this effect and the resulting regret, it is best to preselect the top arms.*

In view of this example, the learner can interpolate between the two extreme cases of users by varying γ : Taking a sufficiently large γ, the learner can model a very precise user for whom it suffices to preselect a subset that just entails the best item. Using a sufficiently small γ instead, the learner is able to reproduce a random user for whom it is best to compose the preselection of the best items.

Table 2. Problem instances with different optimal subsets (indicated in bold font) for the regret in (8) with $n = 5$ and $l = 3$ (omitted subsets had smaller utilities throughout).

S	$\{1, 2, 3\}$	$\{1, 2, 4\}$	$\{1, 2, 5\}$	$\{1, 3, 5\}$	$\{2, 3, 4\}$
$v = (1\,, 0.681\,, 0.572\,, 0.543\,, 0.171), \gamma = 1$					
$\mathrm{U}(S)$	0.795	0.791	**0.806**	0.778	0.605
$v = (1\,, 0.681\,, 0.572\,, 0.543\,, 0.171), \gamma = 20$					
$\mathrm{U}(S)$	**1.000**	**1.000**	**1.000**	**1.000**	0.676
$v = (1\,, 0.681\,, 0.572\,, 0.543\,, 0.171), \gamma = 0.05$					
$\mathrm{U}(S)$	**0.753**	0.744	0.630	0.593	0.599

Remark 3. Note that, in terms of regret, the case of a very precise user is related to the weak regret of the dueling bandits problem (Yue et al., 2012; Chen & Frazier, 2017), where no regret occurs whenever the best arm is participating in the duel. Similarly, the semantics of the regret in the case of a random user corresponds to the strong regret of the dueling bandits problem, where the regret is zero only when the best arm is duelled with itself.

4.3. Lower Bounds

In this section, we prove lower bounds on the expected regret defined in (9) for the two types of Pre-Bandit problems.

Theorem 4.1. *[Restricted Preselection Bandits] Let $n \in \mathbb{N}$, $l \leq n/4$, and $T \geq n$ be integers. Then, for any algorithm φ suggesting an l-sized subset S_t^{φ} at time t,*

$$\begin{aligned} \mathbb{E}_\theta \ \mathcal{R}(T) \ &= \ \mathop{}\limits_{t=1}^{T} \mathbb{E}_\theta \ \mathrm{U}(S^*) - \mathrm{U}(S_t^{\varphi}) \\ &\geq \min\{1, 1/\gamma\}\, C\, \sqrt{n\,T} \end{aligned}$$

holds for any $\theta \in \Theta$ and any $\gamma \in (0, \infty)$, where $C > 0$ is some constant independent of n, l, and T.

Remark 4. The order of the lower bound in Theorem 4.1 coincides with the lower bound on the expected regret derived by Chen & Wang (2018) for the DAS problem under the MNL model with capacity constraints. In particular, the preselection size l does not affect the order, at least if it is smaller than $n/4$. Although the lower bounds are theoretically of the same order, it is not directly possible to use the lower bound results of Agrawal et al. (2016) or Chen & Wang (2018), as in both proofs the probability of the no-choice option is assumed to be strictly positive, and the revenues all equal 1. Moreover, the results for the stochastic click model are quite different from ours (cf. Theorem 2 by Lattimore et al. (2018)), as there the size of the subset l is present in the lower bound. Therefore, we provide a proof in the supplementary material (Section A).

Theorem 4.2. *[Flexible Preselection Bandits] Let $n \in \mathbb{N}$ and $T \geq n$ be integers. Then, for any algorithm φ suggesting subset $S_t^{\varphi} \in \mathbb{A}_{full}$ at time t, the following holds for any $\gamma \in (0, \infty)$:*

(i) [Gap-independent version] There exists a constant $C > 0$ independent of n and T, such that

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta \ \mathcal{R}(T) \ \ \geq C \ \min\{1, 1/\gamma\} \sqrt{T} \, .$$

(ii) [Gap-dependent version] If φ is a no-regret algorithm (cf. Definition 2 in (Saha & Gopalan, 2019b)), there exists a constant $C > 0$ independent of n and T, such that

$$\sup_{\theta \in \Theta} \Big(\min_{i \notin S^*} (\theta_{max} - \theta_i) \ \cdot \ \mathbb{E}_\theta \ \mathcal{R}(T) \ \Big) \\ \geq C \ \min\{1, 1/\gamma\} \, (n-1) \, \log(T) \, .$$

Remark 5. Note that the gap-independent lower bound is independent of the number of arms n. This is in line with the enhancement for the DAS problem for the uncapacitated compared to the capacitated MNL model (Wang et al., 2018). On the other hand, the gap-dependent lower bound depends on the number of arms n, and is of the same order as in the dueling bandit setting. In particular, compared to the dueling bandits setting, there is (theoretically) no improvement by offering subsets larger than two. This is in accordance with the observations made by Saha & Gopalan (2018; 2019b).

5. Algorithms

In this section, we propose the *Thresholding-Random-Confidence-Bound* (TRCB) algorithm stated in Algorithm 1. This algorithm returns subsets $S_1, \ldots, S_T$ for the restricted Pre-Bandit problem. As will be shown, it has a satisfactory upper bound for the expected cumulative regret in (9). For the flexible Pre-Bandit problem, we further suggest the *Confidence-Bound-Racing* (CBR) algorithm as stated in Algorithm 2. It is inspired by the idea of *racing algorithms,* initially introduced by Maron & Moore (1997) to find the best model in the framework of model selection.

5.1. The TRCB Algorithm

First of all, note that an estimation of the score parameter θ is not necessary for the goal of regret minimization. Instead, a proper estimation of the relative scores in (4) is sufficient. Indeed, maximizing the expected utility (6) is equivalent to maximizing the expected utility with respect to some reference arm J, that is

$$\widetilde{\mathrm{U}}(S) = \widetilde{\mathrm{U}}(S; O_J, \gamma) := \frac{{}_{i \in S} \, O_{i,J}^{(1+\gamma)/\gamma}}{{}_{i \in S} \, O_{i,J}} \, , \qquad (10)$$

where $O_J = (O_{1,J}, \ldots, O_{n,J})$, simply because $\widetilde{\mathrm{U}}(S) = v_J^{-1} \cdot \mathrm{U}(S)$.

Thanks to Lemma 1 by Saha & Gopalan (2019a), one can derive appropriate confidence region bounds based on a similar exponential inequality for the relative score estimates, so that one might be tempted to use a UCB-like policy for the restricted Pre-Bandit problem. However, the main problem of such an approach is UCB's principle of "optimism in the face of uncertainty", which tends to exclude arms with low score from a preselection. As we have seen in Example 1, such arms could indeed be part of the optimal subset S^* depending on the specific value of γ.

Algorithm 1 TRCB algorithm

```
input Set of arms $[n]$, preselection size $l \in [2, n] \cap \mathbb{N}$,
    lower bound for score parameters $\theta_{min}$, magnitude of
    uncertainty consideration $C_{shrink} \in (0, 1/2)$, degree
    of preciseness $\gamma$
1: initialization: $W = [w_{i,j}]_{i,j} \leftarrow (0)_{n \times n}$
2: $\hat{O} = [\hat{O}_{i,j}]_{i,j} \leftarrow (1)_{n \times n}$
3: repeat
4:   $t \ \leftarrow t + 1$
5:   $J \leftarrow \arg\max_{i \in [n]} \#\{w_{i,j} \geq w_{j,i} \mid j \neq i\}$
6:   {Break ties arbitrarily}
7:   for $i \in \{1, \ldots, n\} \backslash \{J\}$ do
8:     Sample $\beta_i \sim$ Unif$[\pm\sqrt{\frac{32 \log(l t^{3/2})}{\theta_{min}^4 (w_{i,J} + w_{J,i})}}]$
9:     $\hat{O}_{i,J}^{TRCB} \leftarrow \min \ \theta_{min}^{-1}, \max \ \hat{O}_{i,J} +$
       $C_{shrink} \, \beta_i, \theta_{min}$
10:  end for
11:  Compute $\hat{S} \leftarrow \arg\max_{S \in \mathbb{A}_l} \widetilde{\mathrm{U}}(S; \hat{O}_J^{TRCB}, \gamma)$
12:  Suggest $S_t = \hat{S}$ and obtain choice $i_t \in S_t$
13:  Update $w_{i_t,j} \leftarrow w_{i_t,j} + 1, j \in S_t \backslash \{i_t\}$
     and for $i, j \in S_t$
     $\hat{O}_{i,j} \leftarrow$ $w_{i,j}/w_{j,i},$   $w_{j,i} \neq 0,$
                    $\theta_{min},$   else.
14: until $t == T$
```

The core idea of the TRCB algorithm is to solve this issue with a certain portion of pessimism. Instead of using the upper confidence bound estimates for the relative scores, a random value inside the confidence region of the relative score estimate is drawn (lines 7–8), so that pessimistic guesses for the relative scores are considered as well, which in turn ensures sufficient exploration of the algorithm. This sampling idea can be interpreted as a frequentist statistical version of Thompson Sampling. To exclude inconsistencies with the score parameter space (cf. Section 3.2), these random confidence values are appropriately thresholded.

Until the a priori unknown time horizon is reached (lines 3, 4, 13), the TRCB algorithm repeatedly does the following. Primarily, the arm with the highest total number of wins for the pairwise comparisons is determined as the reference arm J (line 5). Next, for every other arm, a random value inside its confidence region for its relative score with respect to the reference arm J is drawn with uniform distribution and appropriately thresholded (lines 7–10). These thresholded random values correspond to the current belief on the actual relative scores with respect to J and are used to determine the preselection with the highest utility in (10) (line 11). After offering this preselection to the selector and observing its choice (line 12), the pairwise winning counts are updated (by breaking down the l-wise comparison into pairwise comparisons) as well as the estimates for the relative scores (line 13).

The following theorem shows that the upper bound for the worst-case cumulative regret of the proposed TRCB algorithm matches the information-theoretic lower bound on the cumulative regret in Theorem 4.1 with regard to n and T up to a logarithmic term of T (the proof is given in Section B of the supplement).

Theorem 5.1. *If $C_{shrink} \in (0, 1/2)$, then for any $\gamma \in (0, \infty)$ and any $T > n$,*

$$\sup_{\theta\in\Theta} \mathbb{E}_{\theta}^{\mathrm{TRCB}}\, \mathcal{R}(T) \\ \leq C\, \frac{\max\{\theta_{min}^{(\gamma-1)/(\gamma)}, \theta_{min}^{(1-\gamma)/(\gamma)}\}}{\gamma\, \theta_{min}^{2(3+\gamma)}}\, \sqrt{n\, T \log(T)},$$

where $C > 0$ is some constant independent of n, l, T as well as θ_{min} and γ.

Remark 6. The maximization over $\mathbb{A}_l$ in Algorithm 1 (line 11) can be realized by Algorithm 3 provided in the supplementary material. It keeps the computational cost low by exploiting structural properties of the utility function U and the most preferred subsets (see Section 4.2).

5.2. The CBR Algorithm

The CBR algorithm is structurally similar to the TRCB algorithm. However, it uses estimates of the pairwise winning probabilities and the corresponding confidence intervals instead of the relative scores.

In particular, the CBR algorithm maintains a pool of candidates $A \in [n]$ and admits an arm $i \in A$ to be part of the preselection with a certain probability determined by the rate of uncertainty that i could beat the current arm J with the most winning counts. This uncertainty is expressed through the ratio between the length of the confidence interval for $q_{i,J}$ (cf. the definition in (1)) exceeding $1/2$ and the overall confidence interval's length. More specifically, if $[l_i(t), u_i(t)]$ is the confidence interval for $q_{i,J}$ in time instance t, then arm i is included into the preselection with probability $\sigma\ {}^{(u_i(t)-1/2)}\!/\!_{(u_i(t)-l_i(t))}$, where $\sigma : \mathbb{R} \to [0, 1]$ is a sigmoidal function, i.e., a surjective monotone function with $\sigma(1/2) = 1/2$ and $\sigma(x) > 0$ iff $x > 0$. Note that the degree of preciseness γ can be taken into account by the learner through the shape of σ.

Hence, if the confidence interval lies mostly above $1/2$, that is $l_i(t) \approx 1/2$, the chance is high that this particular arm could possibly beat the current best arm and consequently has a large probability of being included in the preselection. In contrast, if the upper bound of the confidence interval is beneath $1/2$, that is $u_i(t) \leq 1/2$, the arm is discarded from the pool of candidates (lines 12–14), as one can be sure that this arm is already beaten by another.

Algorithm 2 CBR-algorithm

```
input  Set of arms $[n]$, sigmoidal function $\sigma : \mathbb{R} \to [0, 1]$
 1: Initialization: $W = [w_{i,j}] \leftarrow (0)_{n\times n}$
 2: $\hat{Q} = [\hat{q}_{i,j}] \leftarrow ({}^1\!/\!_2)_{n\times n},\ A \leftarrow [n]$
 3: repeat
 4:    $t \leftarrow t + 1$
 5:    $J \leftarrow \arg\max_{i\in[n]} \#\{w_{i,j} \geq w_{j,i} \mid j = i\}$
 6:    {Break ties arbitrarily}
 7:    $S \leftarrow \{J\}$
 8:    for $i \in A$ do
 9:       $c_i \leftarrow \sqrt{{}^{2\log(nt^{3/2})}\!/\!_{(w_{i,J}+w_{J,i})}}$
10:       $\hat{t}_{i,J} \leftarrow \sigma\ {}^{(\hat{q}_{i,J}+c_i-1/2)}\!/\!_{2c_i}$
11:       $S \leftarrow \begin{matrix} S \cup \{i\}, & \text{with probability } \hat{t}_{i,J} \\ S, & \text{with probability } 1 - \hat{t}_{i,J} \end{matrix}$
12:       if $\hat{t}_{i,J} = 0$ then
13:          $A \leftarrow A \backslash \{i\}$
14:       end if
15:    end for
16:    Suggest $S_t = S$ and obtain choice $i_t \in S_t$
17:    Update $w_{i_t,j} \leftarrow w_{i_t,j} + 1, j \in S_t \backslash \{i_t\}$
       and $\hat{q}_{i,j} \leftarrow {}^{w_{i,j}}\!/\!_{(w_{i,j}+w_{j,i})}$ for $i, j \in S_t$
18: until $t == T$
```

At the beginning, the major part of the arms have a high chance to be part of the preselection, which however decreases over the course of time until finally the preselection consists of only the best arm(s). In the repetition phase, the preselection is successively built starting from the current arm with the most total number of wins for the pairwise comparisons and adding arms from the active set depending on the outcome of a Bernoulli experiment (lines 5–11), whose success probability depends on the length of the confidence interval (of the arm's pairwise winning probability against J) above $1/2$. After offering this preselection to the selector and observing its choice (line 16), the pairwise winning counts and estimates on the pairwise winning probabilities are updated (line 17).

We have the following theorem for the upper bound on the cumulative regret for CBR, which matches the information-theoretic gap-dependent lower bound on the cumulative regret in Theorem 4.2 (the proof is given in Section C of the supplement).

Theorem 5.2. *There are universal constants $C_0, C_1 > 0$,*

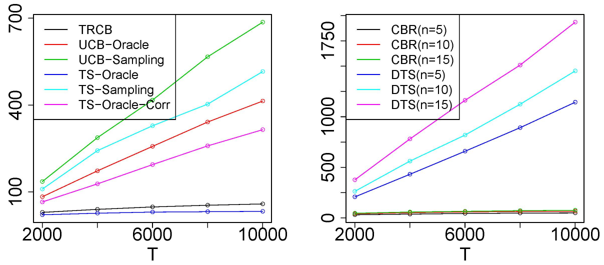

Figure 1. Left: Mean cumulative regret for 1000 runs of randomly generated restricted Pre-Bandit instances. Right: Mean cumulative regret for 1000 runs of randomly generated flexible Pre-Bandit instances.

which do not dependent on T or n, such that

$$\sup_{\theta\in\Theta}\mathbb{E}_{\theta}^{\mathrm{CBR}}\,\mathcal{R}(T) \leq C_0\, n + C_1\,\frac{\max\{\theta_{min}^{(\gamma-1)/(\gamma)}, \theta_{min}^{(1-\gamma)/(\gamma)}\}}{\gamma}\times {}_{i\in[n]\setminus S^*}\frac{\log(T)}{\theta_{max}-\theta_i}$$

for any $T > n$, $\gamma \in (0,\infty)$.

6. Experiments

In this section, we investigate the performance of TRCB (Algorithm 1) as well as CBR (Algorithm 2) on synthetic data for some specific scenarios, while providing further scenarios in the supplementary material.

6.1. Restricted Pre-Bandit Problem

First, we analyze the empirical regret growth with varying time horizon T for the restricted Pre-Bandit problem. We consider the case $n = 10$, $l = 3$, and time horizons $T \in \{i \cdot 2000\}_{i=1}^{5}$. The degree of preciseness is $\gamma = 1$ throughout, and the score parameters $\theta = (\theta_i)_{i\in[n]}$ are drawn uniformly at random from the n-simplex, i.e., without a restriction on their minimal value and thus allowing θ_{min} to be infinitesimal. The left plot in Figure 1 provides the performance of our algorithms together with some algorithms for the DAS problem (see Section E in the supplement for more information on these).

For the algorithms of the DAS problem, the best arm is set to be the no-choice option, thereby putting (most of) them in the advantageous position of knowing a priori one element of the optimal subset. Nevertheless, only TS-Oracle, with the advantage of knowing the best arm a priori, is able to slightly outperform TRCB in this scenario, whereas all other algorithms are distinctly outperformed by TRCB.

To explain this observation, recall our remark on UCB-like strategies in Section 5.1. The UCB-based algorithms UCB-Oracle resp. UCB-Sampling as well as the UCB-like approximation of the variance of TS-Oracle-Corr tend to exclude arms with a low score from the suggested subset, even though they are contained in the optimal preselection. TS-Oracle and TS-Sampling, which do not use upper confidence bounds and include low score arms in the suggested subsets, are performing much better. The gap between these two TS algorithms shows how heavily the algorithms depend on the assumption that the no-choice option corresponds to the highest scored arm, since we designed TS-Sampling such that, in each run, it samples once the best-arm from the top three arms according to an MNL model.

In summary, this simulation confirms that the introduced (restricted) Pre-Bandit problem is indeed a new framework that differs from the DAS problem. A naïve application of existing methods for the DAS problem is not suitable for this kind of problem.

6.2. Flexible Pre-Bandit Problem

Next, we investigate the empirical regret growth with varying time horizon T and varying numbers of arms n for the flexible Pre-Bandit problem. In addition, we compared our algorithms with the Double Thompson Sampling (DTS) algorithm by Wu & Liu (2016), which is considered state-of-the-art for the dueling bandits problem with a small numbers of arms Sui et al. (2017).

In the right picture of Figure 1, the results are displayed for the CBR resp. DTS algorithm on 1000 repetitions, respectively, with $n \in \{5, 10, 15\}$, $T \in \{i \cdot 2000\}_{i=1}^{5}$, and $\sigma(x) = (1 \wedge x)1_{[0,\infty)}(x)$. The score parameters are generated randomly as before. It is clearly recognizable that CBR distinctly outperforms DTS in all scenarios, indicating that offering larger subsets is at least experimentally beneficial to find the best arm more quickly.

7. Conclusion

In this paper, we have introduced the Pre-Bandit problem as a practically motivated and theoretically challenging variant of preference-based multi-armed bandits in a regret minimization setting. More specifically, we proposed two scenarios, one in which preselections are of fixed size and another one in which the size is under the control of the agent. For both scenarios, we derived lower bounds on the regret of algorithms solving these problems. Moreover, we proposed concrete algorithms and analyzed their performance theoretically and experimentally.

Our new framework suggests a multitude of conceivable paths for future work. Most naturally, it would be interesting to analyze the Pre-Bandit problem under different assumptions on the user's choice behavior—despite being natural and theoretically justified (McFadden, 2001; Train, 2009), the assumption of the PL model is relatively strong, and the question is to what extent it could be relaxed. The main challenge surely lies in defining a sensible notion of regret, but an extension to the nested logit-model (Chen et al., 2018c) or considering contextual information (Chen et al., 2018b) seems to be possible. However, it is worth noting that our derived lower bounds on the regret based on expected utilities in the spirit of (6) even hold for more general choice models, such as generalized random utility models (Walker & Ben-Akiva, 2002; Train, 2009), which encompass the PL model.

Last but not least, like the related dynamic assortment selection problem studied in operational research, the motivation of our new framework stems from practical applications. Therefore, we are also interested in applying our algorithms to real-world problems, such as algorithm (pre-)selection already mentioned in the introduction. In particular, the realm of algorithm selection seems to be a canonical candidate for our setting, as the decisions made are based on the noisy performance values of the algorithms, which justify a stochastic modeling of the process.

ACKNOWLEDGMENTS

The authors gratefully acknowledge financial support by the Germany Research Foundation (DFG). Moreover, the authors would like to thank the Paderborn Center for Parallel Computation (PC2) for the use of the OCuLUS cluster.

References

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. A near-optimal exploration-exploitation approach for assortment selection. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pp. 599–600, 2016.

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. Thompson sampling for the mnl-bandit. In *Conference on Learning Theory*, pp. 76–78, 2017.

Alvo, M. and Yu, P. L. *Statistical methods for ranking data*. Springer, 2014.

Busa-Fekete, R., Hüllermeier, E., and Mesaoudi-Paul, A. E. Preference-based online learning with dueling bandits: A survey. *arXiv preprint arXiv:1807.11398*, 2018.

Caro, F. and Gallien, J. Dynamic assortment with demand learning for seasonal consumer goods. *Management Science*, 53(2):276–292, 2007.

Cesa-Bianchi, N. and Lugosi, G. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

Chen, B. and Frazier, P. I. Dueling bandits with weak regret. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 731–739, 2017.

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

Chen, X., Li, Y., and Mao, J. 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*, pp. 2504–2522, 2018a.

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

Chen, X., Wang, Y., and Zhou, Y. Dynamic assortment selection under the nested logit models. *arXiv preprint arXiv:1806.10410*, 2018c.

Dimara, E., Bezerianos, A., and Dragicevic, P. The attraction effect in information visualization. *IEEE transactions on visualization and computer graphics*, 23(1): 471–480, 2017.

Kerschke, P., Hoos, H. H., Neumann, F., and Trautmann, H. Automated algorithm selection: Survey and perspectives. *Evolutionary computation*, 27(1):3–45, 2019.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pp. 535–543, 2015.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press (to appear), 2020.

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

Luce, R. D. Individual choice behavior: a theoretical analysis. *Wiley*, 1959.

Maron, O. and Moore, A. W. The racing algorithm: Model selection for lazy learners. *Artificial Intelligence Review*, 11(1-5):193–225, 1997.

McFadden, D. Economic choices. *American economic review*, 91(3):351–378, 2001.

Plackett, R. The analysis of permutations. *Applied Statistics*, pp. 193–202, 1975.

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

Saha, A. and Gopalan, A. Battle of bandits. In *Uncertainty in Artificial Intelligence*, 2018.

Saha, A. and Gopalan, A. Pac battling bandits in the plackett-luce model. In *Algorithmic Learning Theory*, pp. 700–737, 2019a.

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

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

Sui, Y., Zhuang, V., Burdick, J. W., and Yue, Y. Multi-dueling bandits with dependent arms. In *Uncertainty in Artificial Intelligence*, 2017.

Train, K. E. *Discrete choice methods with simulation*. Cambridge university press, 2009.

Walker, J. and Ben-Akiva, M. Generalized random utility model. *Mathematical social sciences*, 43(3):303–343, 2002.

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

Wu, H. and Liu, X. Double thompson sampling for dueling bandits. In *Advances in Neural Information Processing Systems*, pp. 649–657, 2016.

Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

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