Simultaneously Learning Stochastic and Adversarial Bandits under the Position-Based Model

Cheng Chen[1], **Canzhe Zhao**[2], **Shuai Li** [2*]

[1] Nanyang Technological University
[2] Shanghai Jiao Tong University
cheng.chen@ntu.edu.sg, canzhezhao@sjtu.edu.cn, shuaili8@sjtu.edu.cn

Abstract

Online learning to rank (OLTR) interactively learns to choose lists of items from a large collection based on certain click models that describe users' click behaviors. Most recent works for this problem focus on the stochastic environment where the item attractiveness is assumed to be invariant during the learning process. In many real-world scenarios, however, the environment could be dynamic or even arbitrarily changing. This work studies the OLTR problem in both stochastic and adversarial environments under the position-based model (PBM). We propose a method based on the follow-the-regularized-leader (FTRL) framework with Tsallis entropy and develop a new self-bounding constraint especially designed for PBM. We prove the proposed algorithm simultaneously achieves $O(\log T)$ regret in the stochastic environment and $O(m\sqrt{nT})$ regret in the adversarial environment, where T is the number of rounds, n is the number of items and m is the number of positions. We also provide a lower bound of order $\Omega(m\sqrt{nT})$ for adversarial PBM, which matches our upper bound and improves over the state-of-the-art lower bound. The experiments show that our algorithm could simultaneously learn in both stochastic and adversarial environments and is competitive compared to existing methods that are designed for a single environment.

Introduction

Learning to rank is widely used in online web search and recommender systems which selects a small group of items to present in a limited number of positions after a user starts a search session (Liu et al. 2009). Online learning to rank (OLTR) is to learn the best ranking policy through user interactions and aims to maximize user satisfaction, *e.g.* the number of user clicks, during the learning period. To understand the click signals received from users on given ranked lists of items, many click models are introduced and studied (Chuklin, Markov, and Rijke 2015). One of the most popular click models adopted in the industry is the position-based model (PBM) (Richardson, Dominowska, and Ragno 2007) due to its simplicity and effectiveness to characterize the click rate as a product of item attractiveness and position bias. PBM is studied in OLTR setting with theoretical analysis on regret (Lagrée, Vernade, and Cappé 2016; Komiyama, Honda, and Takeda 2017), which is in expectation the difference of the received clicks from the clicks of the best policy. Some other works in OLTR study the cascade model (Kveton et al. 2015a; Li et al. 2016; Zong et al. 2016) and general click model (Zoghi et al. 2017; Lattimore et al. 2018; Li, Lattimore, and Szepesvári 2019).

Most existing works in OLTR focus on the stochastic environment where the item attractiveness and position examination probabilities, if any, are assumed to be fixed through the learning process. However, this usually is a strong assumption in real applications where the item attractiveness could change dynamically, like the clothes interest of users might periodically change across seasons. The algorithms designed in the stochastic environment might fail to converge if the stochastic assumptions do not hold. This motivates the study of adversarial environment where the involved samples are arbitrary on a bounded domain. Usually the regret guarantee of algorithms designed under adversarial environment can only be $O(\sqrt{T})$, even in the stochastic environment whose best algorithm can achieve a much better regret of $O(\log(T))$. It is an interesting topic in online learning if there is an algorithm that can achieve $O(\sqrt{T})$ regret if run in the adversarial environment and $O(\log(T))$ regret if run in the stochastic environment. This problem is also called best-of-both-worlds (BOBW). Some works study this problem in classical multi-armed bandit problem (MAB) (Bubeck and Slivkins 2012; Seldin and Slivkins 2014; Auer and Chiang 2016; Seldin and Lugosi 2017; Wei and Luo 2018; Zimmert and Seldin 2019) and combinatorial MAB (CMAB) with semi-bandit feedback (Zimmert, Luo, and Wei 2019). It is an open question if we can design BOBW algorithms in OLTR whose adversarial formulation needs to be well deliberated first. In this work we hope to answer this question under the commonly adopted PBM.

We propose an algorithm for OLTR under PBM based on the follow-the-regularized-leader (FTRL) framework to simultaneously learn in stochastic and adversarial environments. Though OLTR under PBM can be formulated as a special case of CMAB by regarding the pair of an item and a position as a base arm, the direct application of existing studies does not hold. One of the main challenges is that the commonly defined suboptimality gap could be negative in the PBM setting, making it impossible to follow the existing *self-bounding* technique in (Zimmert, Luo, and Wei 2019;

*Corresponding author

Zimmert and Seldin 2019; Wei and Luo 2018). For this, we deliberately design a suboptimality gap, which is non-trival and quite different from the commonly defined suboptimality gap. We also build a new form of self-bounding constraint for PBM based on the property of the proposed suboptimality gap. Also, the structure of PBM could have Tsallis entropy as the potential function and the corresponding regularized leader can be computed efficiently, compared with the hybrid regularizer in the previous work (Zimmert, Luo, and Wei 2019). We prove our algorithm could achieve $O(\log T)$ regret in the stochastic environment and $O(\sqrt{T})$ regret in the adversarial environment, verifying its simultaneous learning ability in both environments.

Furthermore, we provide a regret lower bound $\Omega(m\sqrt{nT})$ for OLTR under PBM. This improves the state-of-the-art lower bound $\Omega(\sqrt{mnT})$ (Lattimore et al. 2018) which is analyzed under document-based model, a special case of PBM. Our lower bound matches our upper bound, showing the optimality of both. Table 1 shows a full comparison of our work with most related works.

The experiments show that our algorithm outperforms the baselines in adversarial environments while is competitive with `TopRank` (Lattimore et al. 2018) and `PMED` (Komiyama, Honda, and Takeda 2017) in stochastic environments. The results show the simultaneous learning ability of our algorithm in both environments.

Related Work The study of OLTR under PBM has received many interests. For the stochastic environment, Lagrée, Vernade, and Cappé (2016) studies PBM but assumes the position examination probabilities are known or could be pre-computed from historical data. This assumption is a bit unrealistic and does not account for possible drift of position bias. Komiyama, Honda, and Takeda (2017) remove this requirement but only provides an asymptotic regret bound. With rank-1 structure, PBM with unknown position bias can also be solved using methods in rank-1 bandits (Katariya et al. 2017) though their setting is originally designed to select one item each round. Some works study a general class of click models with PBM as a special case (Zoghi et al. 2017; Lattimore et al. 2018; Li, Lattimore, and Szepesvári 2019). They distill a set of assumptions that are satisfied by common click models including the cascade model and PBM. The algorithms designed on this general click model are more robust than that on PBM. All the above algorithms study only the stochastic environment and might be brittle when the stochastic assumption is violated.

For the adversarial environment, PBM is first studied by Kale, Reyzin, and Schapire (2010) as an ordered slate model. They solve it by a variant of multiplicative-weights algorithm and prove a regret upper bound $O(m\sqrt{n\log(n)T})$, $O(\sqrt{\log(n)})$ worse than ours. Bubeck and Cesa-Bianchi (2012) show that OSMD with a 0-potential function can achieve $O(m\sqrt{nT})$ regret, but their method need to know the time horizon. Radlinski, Kleinberg, and Joachims (2008) study a ranked bandit problem using the greedy idea to select items one-by-one, which can only give approximation guarantees. It is extended to metric space by considering item contexts (Slivkins, Radlinski, and Gollapudi 2013). Other studies include online optimization over the permutahedron (Ailon 2014; Ailon, Hatano, and Takimoto 2016), which corresponds to PBM with $m = n$ and PBM with full-information feedback (Cohen and Hazan 2015).

The model of OLTR under PBM can be regarded as a special case of combinatorial semi-bandits (Gai, Krishnamachari, and Jain 2012; Chen, Wang, and Yuan 2013; Kveton et al. 2015b; Combes et al. 2015; Combes, Magureanu, and Proutiere 2017; Zimmert, Luo, and Wei 2019; Neu and Bartók 2013; Neu 2015; Audibert, Bubeck, and Lugosi 2014) with specific combinatorial constraints. Most existing works study either stochastic or adversarial environment.

For the BOBW algorithms, many study this topic for MAB (Bubeck and Slivkins 2012; Seldin and Slivkins 2014; Auer and Chiang 2016; Seldin and Lugosi 2017; Wei and Luo 2018) where Zimmert and Seldin (2019) show that FTRL with $\frac{1}{2}$-Tsallis entropy can achieve optimal regret bounds for both stochastic and adversarial environments. Zimmert, Luo, and Wei (2019) study combinatorial semi-bandits by a novel hybrid regularizer but only show the optimality in two special cases for the stochastic environment, full combinatorial set and m-set. Other BOBW works include prediction with expert advice (Koolen, Grünwald, and van Erven 2016; Mourtada and Gaïffas 2019), linear bandits (Lee et al. 2021), online convex optimization (Cutkosky and Boahen 2017) and Markov decision process (Jin and Luo 2020). Our work focuses on the BOBW under PBM.

Setting

This section introduces both stochastic and adversarial environments of OLTR under PBM.

Suppose there are n items with item set $[n] = \{1, 2, \dots, n\}$ and m positions ($m \le n$). In each round t, the learner selects an ordered list $I_t = (i_{t,1}, i_{t,2}, \dots, i_{t,m})$ consisting of m distinct items, where $i_{t,j}$ denotes the item placed at position j in round t. Note that this problem can be formulated as a special case of combinatorial semi-bandits and the action list I_t can be written as a subpermutation matrix $X_t \in \mathcal{X}$ where

$$\mathcal{X} = \left\{ X \in \{0,1\}^{n\times m} \;\middle|\; \sum_{i=1}^{n} X_{i,j} = 1, \forall j \in [m]; \right. \\ \left. \sum_{j=1}^{m} X_{i,j} \le 1, \forall i \in [n] \right\}$$

is the action set and $X_{i,j}$ denotes whether to put item i on position j.

Lattimore and Szepesvári (2020) introduce the adversarial setting of PBM as follows. For each round t and position $k \in [m]$, the environment secretly chooses $S_{t,k}$ as subset of $[n]$. The reward of round t is defined as $r_t = \sum_{k=1}^{m} \mathbb{1}\{i_{t,k} \in S_{t,k}\}$ where I_t is the selected action list at time t. The feedback is the positions of the clicked items. Notice that this model can be reformulated as a combinatorial semi-bandit problem. At round t, the environment secretly chooses all loss matrices $\ell_t \in \{0,1\}^{n\times m}$ for any t

	Regret Bound (Stochastic)	Regret Bound (Adversarial)	Original Model
Kale, Reyzin, and Schapire (2010)	-	$O\left(m\sqrt{nT\log(n)}\right)$	Bandits for Ordered Slates
Bubeck and Cesa-Bianchi (2012)	-	$O\left(m\sqrt{nT}\right)$	CMAB with semi-bandit feedback
Lagrée, Vernade, and Cappé (2016)	$O\left(\frac{n}{\beta_m\Delta}\log(T)\right)$	-	PBM with known position bias
Zoghi et al. (2017)	$O\left(\frac{m^3n}{\Delta}\log(T)\right)$	-	General Click Model
Lattimore et al. (2018)	$O\left(\frac{mn}{\Delta}\log(T)\right)$ $O\left(\sqrt{m^3nT\log(T)}\right)$ $\Omega\left(\sqrt{mnT}\right)$	$\Omega\left(\sqrt{mnT}\right)$	General Click Model
Li, Lattimore, and Szepesvári (2019)	$O\left(m\sqrt{nT\log(nT)}\right)$	-	General Click Model with Linear Features
Zimmert, Luo, and Wei (2019)	$O\left(\frac{m^2n}{\Delta_\beta\Delta}\log(T)\right)$	$O\left(m\sqrt{nT}\right)$	CMAB with semi-bandit feedback
Ours	$O\left(\frac{mn}{\Delta_\beta\Delta}\log(T)\right)$	$O\left(m\sqrt{nT}\right)$ $\Omega\left(m\sqrt{nT}\right)$	

Table 1: This table compares regret bounds of related works when their results are applied to OLTR under PBM for both stochastic and adversarial environments. T is the number of total rounds, m is the number of positions and n is the number of items. Δ is the minimal gap between the attractiveness of the best m items. Δ_β is the minimal gap between position examination probabilities.

before the game where $\ell_{t,i,j} = 0$ means there is no loss (or there is a click) if placing item i at position j in round t. After selecting I_t, the algorithm receives a loss of $\langle X_t, \ell_t\rangle$ and observes semi-bandit feedback $\ell_{t,i,j}$ for those (i,j) such that $X_{t,i,j} = 1$. The goal of the algorithm is to minimize the expected cumulative pseudo-regret

$$R(T) = \mathbb{E}\left[\sum_{t=1}^{T}\langle X_t - x^*, \ell_t\rangle\right], \qquad (1)$$

where $x^* \in \arg\min_{x\in\mathcal{X}} \mathbb{E}\left[\sum_{t=1}^{T}\langle x, \ell_t\rangle\right]$ is the best action and the expectation is taken over the randomness of both the algorithm and the environment.

For the stochastic environment, each item $i \in [n]$ is associated with an (unknown) attractiveness $\alpha_i \in [0,1]$ and each position $j \in [m]$ is associated with an (unknown) examination probability $\beta_j \in (0,1]$. Without loss of generality, we assume $\alpha_1 > \alpha_2 > \cdots > \alpha_m > \alpha_{m+1} \geq \alpha_{m+2} \geq \cdots \geq \alpha_n$ and $\beta_1 > \beta_2 > \cdots > \beta_m > \beta_{m+1} = 0$. Let $\mathcal{H}_t$ be the σ-algebra containing all the history $(\ell_1, X_1, \ldots, \ell_t, X_t)$ by the end of round t. In the stochastic environment, all elements in the loss matrix ℓ_t are $\mathcal{H}_{t-1}$-conditionally independent whose (i,j)-entry is drawn from Bernoulli distribution $\mathrm{Ber}(1-\alpha_i\beta_j)$ like previous works (Komiyama, Honda, and Takeda 2017; Lagrée, Vernade, and Cappé 2016; Chuklin, Markov, and Rijke 2015). In this case, $x^*_{i,j}$ is actually $\delta_{i,j}$ which is 1 if and only if $i = j$. The goal of the algorithm is also to minimize the expected cumulative pseudo-regret Eq.(1).

Notations Throughout this paper, we use I_j^* to denote the item selected by the best action x^* at position j, or $x^*_{I_j^*,j} = 1$. For a given set $\mathcal{X}$, let $\mathbb{1}\{\mathcal{X}\}$ be the indicator function and $\mathcal{I}_{\mathcal{X}}(x)$ be the characteristic function which is ∞ if $x \notin \mathcal{X}$ and 0 otherwise. Let $\mathrm{Conv}(\mathcal{X})$ be the convex hull of $\mathcal{X}$. We use $\mathbf{1}_n$ to denote the n-dimensional vector whose entries are all 1s. The conditional expectation $\mathbb{E}\left[\cdot \mid \mathcal{H}_{t-1}\right]$ is abbreviated as $\mathbb{E}_t\left[\cdot\right]$. For the stochastic environment, let $\Delta = \min_{i\in[m]}\{\alpha_i - \alpha_{i+1}\}$ be the minimal gap between the attractiveness of top m items and $\Delta_\beta = \min_{j\in[m]}\{\beta_j - \beta_{j+1}\}$ be the minimal gap between any two position examination probabilities.

Algorithm

This section presents our main algorithm, `FTRL-PBM`, in Algorithm 1 for both stochastic and adversarial environments under PBM. Our `FTRL-PBM` algorithm follows the general follow-the-regularized leader (FTRL) framework, whose main idea is to follow the action which minimizes the regularized cumulative loss of the past rounds. Since the

Algorithm 1: FTRL-PBM

Input: $\mathcal{X}$.

1: $\hat{L}_0 = \mathbf{0}_{n\times m}, \eta_t = 1/(2\sqrt{t})$.
2: **for** $t = 1, \dots, T$ **do**
3: Compute

$$x_t = \underset{x\in\mathrm{Conv}(\mathcal{X})}{\arg\min} \langle x, \hat{L}_{t-1}\rangle + \frac{1}{\eta_t}\Psi(x)\,;$$

4: Sample $X_t \sim P(x_t)$;
5: Observe $\ell_{t,i,j}$ for those (i,j)-th entries satisfying $X_{t,i,j} = 1$;
6: Compute the loss estimator $\hat{\ell}_{t,i,j} = \frac{\ell_{t,i,j}\cdot\mathbb{1}\{X_{t,i,j}=1\}}{x_{t,i,j}}$;
7: Compute $\hat{L}_t = \hat{L}_{t-1} + \hat{\ell}_t$.
8: **end for**

complete loss vectors cannot be observed in the bandit setting, usually an unbiased estimator $\hat{\ell}_t$ satisfying $\mathbb{E}_t\left[\hat{\ell}_t\right] = \ell_t$ would serve as a surrogate.

Specifically, our algorithm FTRL-PBM keeps track of the cumulative estimated loss $\hat{L}_t = \sum_{s=1}^t \hat{\ell}_s \in \mathbb{R}_+^{n\times m}$ and initializes it as a zero vector (line 1). At each round t, FTRL-PBM first computes a regularized leader x_t in the convex hull of the action set $\mathrm{Conv}(\mathcal{X})$ by

$$x_t = \underset{x\in\mathrm{Conv}(\mathcal{X})}{\arg\min} \langle x, \hat{L}_{t-1}\rangle + \frac{1}{\eta_t}\Psi(x) \tag{2}$$

where $\Psi(x)$ is the regularizer (line 3). Here we take the $1/2$-Tsallis entropy

$$\Psi(x) = \sum_i -\sqrt{x_i}$$

as our regularizer, which is shown optimal for BOBW MAB (Zimmert and Seldin 2019). For BOBW semi-bandits, the optimal algorithm adopts a hybrid regularizer (Zimmert, Luo, and Wei 2019), which may not be efficient for PBM and will be discussed later.

Then FTRL-PBM samples an action $X_t \sim P(x_t)$ from $\mathcal{X}$ (line 4) where $P(x_t)$ satisfies $\mathbb{E}_{X\sim P(x_t)}[X] = x_t$. We follow previous works (Kale, Reyzin, and Schapire 2010; Helmbold and K Warmuth 2009) to construct $P(x)$. The method is to first complete matrix x_t into a doubly stochastic matrix M_t, which is a convex combination of permutation matrices by Birkhoff's theorem, and then decompose matrix M_t into its convex combination of at most n^2 permutation matrices by Algorithm 1 of (Helmbold and K Warmuth 2009). The time complexity of whole sampling procedure is of order $O(n^{4.5})$ and the details can be found in Appendix.

After observing the semi-bandit feedback for the selected action X_t, we can construct the unbiased estimator for the loss vector as

$$\hat{\ell}_{t,i,j} = \frac{\ell_{t,i,j}\cdot\mathbb{1}\{X_{t,i,j}=1\}}{x_{t,i,j}}$$

for the (i,j)-th entry (line 6). Then the cumulative estimated loss $\hat{L}$ is updated (line 7).

Optimization

It remains to solve the optimization problem Eq.(2). Let $f(x) = \langle x, \mathbf{1}\rangle + \Psi(x)$. Then Eq.(2) is equivalent to the following problem

$$x_t = \underset{x\in\mathrm{Conv}(\mathcal{X})}{\arg\min} \langle x, \hat{L}_{t-1}\rangle + \frac{1}{\eta_t}f(x)\,.$$

Since $f(x)$ is a Legendre function and $\mathrm{Conv}(\mathcal{X})$ is compact, the above equation can be solved by the following two-step procedure (see Chapter 28 of (Lattimore and Szepesvári 2020))

$$\tilde{x}_t = \arg\min \langle x, \hat{L}_{t-1}\rangle + \eta_t^{-1}f(x)\,, \tag{3}$$

$$x_t = \underset{x\in\mathrm{Conv}(\mathcal{X})}{\arg\min} D_f(x, \tilde{x}_t)\,, \tag{4}$$

where D_f denotes the Bregman divergence induced by the Legendre function f. Since $\hat{L}_{t-1}$ is non-negative, we can guarantee that the minimum of Eq.(3) exists. Note that the linear term in $f(x)$ is necessary because $\langle x, \hat{L}_{t-1}\rangle + \frac{1}{\eta_t}\Psi(x)$ does not have a minimum for the unconstrained case when any element of $\hat{L}_{t-1}$ is equal to zero.

However, solving Eq.(4) is still not easy. Our idea is to leverage the cyclic Bregman projection (CBP) (Bregman 1967; Bauschke and Lewis 2000). We first decompose $\mathrm{Conv}(\mathcal{X})$ into two sets on which Bregman projection can be efficiently computed. Note that (Dunn and Harshbarger 1978; Mendelsohn and Dulmage 1958)

$$\mathrm{Conv}(\mathcal{X}) = \left\{ X\in[0,1]^{n\times m} \,\middle|\, \sum_{i=1}^n X_{i,j} = 1, \forall j\in[m]; \sum_{j=1}^m X_{i,j} \le 1, \forall i\in[n] \right\}.$$

Thus there is $\mathrm{Conv}(\mathcal{X}) = \mathcal{X}_1 \cap \mathcal{X}_2$ where

$$\mathcal{X}_1 = \left\{ X\in[0,1]^{n\times m} \,\middle|\, \sum_{i=1}^n X_{i,j} = 1, \forall j\in[m] \right\},$$

$$\mathcal{X}_2 = \left\{ X\in[0,1]^{n\times m} \,\middle|\, \sum_{j=1}^m X_{i,j} \le 1, \forall i\in[n] \right\}.$$

Then we can solve Eq.(4) by the CBP algorithm presented in Algorithm 2. The Bregman projection onto $\mathcal{X}_1$ can be computed efficiently. Since the constraint in $\mathcal{X}_1$ is column-wise, we can compute the projection of each column separately. For each column j, we need to solve

$$\underset{\substack{\sum_{i=1}^n x_{i,j}=1\\ x_{i,j}\in[0,1]}}{\arg\min} -\sum_{i=1}^n \sqrt{x_{i,j}} + \langle \mathbf{1} - \nabla f(\tilde{x}_t)_j, x_j\rangle$$

where $\nabla f(\tilde{x}_t)_j$ is the j-th column of $\nabla f(\tilde{x}_t)$. Similar to the implementation of (Zimmert and Seldin 2019), its solution takes the form

$$x_{i,j} = \frac{1}{4}(\nabla f(\tilde{x}_t)_{i,j} - \lambda - 1)^{-2} \tag{5}$$

Algorithm 2: Cyclic Bregman Projection

1: $x^0 = \tilde{x}_t, s = 0$.
2: **while** not converge **do**
3: $s = s + 1$
4: $x^{2s-1} = \arg\min_{x \in \mathcal{X}_1} D_f(x, x^{2s-2})$
5: $x^{2s} = \arg\min_{x \in \mathcal{X}_2} D_f(x, x^{2s-1})$
6: **end while**

where λ can be solved by

$$\sum_{i=1}^{n} \frac{1}{4} (\nabla f(\tilde{x}_t)_{i,j} - \lambda - 1)^{-2} = 1 \,. \tag{6}$$

This can be figured out approximately using Newton's method with only a few iterations. The Bregman projection onto $\mathcal{X}_2$ is similar and all details are postponed to Appendix.

Remark 1. *Note that the hybrid regularizer adopted in (Zimmert, Luo, and Wei 2019) may not be efficiently solved with low computational complexity by traditional convex optimization methods under PBM. The reason is that the computation of the projection onto the constrained set* $\mathrm{Conv}(\mathcal{X})$ *is rather expensive and the gradient of the hybrid regularizer is not Lipschitz continuous. The Tsallis entropy works since we can get a closed form of Eq.*(5)*. It is an open question whether the FTRL with the hybrid regularizer can be solved efficiently under PBM.*

Regret Analysis

This section provides regret upper bounds of our algorithm `FTRL-PBM` for both stochastic and adversarial environments, together with an improved lower bound for PBM, which also matches our upper bound. We also discuss the relationship between our results and previous works.

Upper Bounds

We give the regret upper bounds of `FTRL-PBM` for each of the adversarial and stochastic environments and provide proof sketches.

Theorem 1. *For the adversarial environment, the regret of our* `FTRL-PBM` *algorithm satisfies*

$$R(T) \le 3m + 2m \log T + \sum_{t=1}^{T} \left(\frac{3}{\sqrt{t}} \sum_{j=1}^{m} \sum_{i \neq I_j^*} \sqrt{\mathbb{E}[x_{t,i,j}]} \right) = O(m\sqrt{nT}) \,.$$

Though this regret bound matches that of OSMD with 0-potential (Bubeck and Cesa-Bianchi 2012) and BOBW semi-bandits (Zimmert, Luo, and Wei 2019), these two methods have some shortcomings compared to ours. OSMD with 0-potential needs to know the time horizon. Existing doubling trick methods lead to additional logarithmic factors in either stochastic or adversarial setting (Besson and Kaufmann 2018). BOBW semi-bandits could be inefficient under PBM since they use a hybrid regularizer.

Proof sketch. Denote $\Psi_t(\cdot) = \frac{1}{\eta_t}\Psi(\cdot)$. Let $\Phi_t(\cdot) = \max_{x \in \mathrm{Conv}(\mathcal{X})} \langle x, \cdot \rangle - \Psi_t(x)$ be the Fenchel conjugate of $\Psi_t + \mathcal{I}_{\mathrm{Conv}(\mathcal{X})}$. Like the standard FTRL analysis (Chapter 28 of (Lattimore and Szepesvári 2020)), the regret can be decomposed as a sum of the stability term and the regularization penalty term

$$R(T) = \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \langle X_t, \ell_t \rangle + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1})\right]}_{R_{stab}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} -\Phi_t(-\hat{L}_t) + \Phi_t(-\hat{L}_{t-1}) - \langle x^*, \ell_t \rangle\right]}_{R_{pen}} .$$

Then we bound these two terms separately (which proofs are postponed to Appendix)

$$R_{stab} \le 3m + 2m \log T + \sum_{t=4}^{T} \left[\frac{1}{\sqrt{t}} \sum_{j=1}^{m} \sum_{i \neq I_j^*} \left(\sqrt{\mathbb{E}[x_{t,i,j}]} + \mathbb{E}[x_{t,i,j}] \right) \right] , \tag{7}$$

$$R_{pen} \le \sum_{t=1}^{T} \sum_{j=1}^{m} \sum_{i \neq I_j^*} \frac{1}{\sqrt{t}} \left(2\sqrt{\mathbb{E}[x_{t,i,j}]} - \mathbb{E}[x_{t,i,j}] \right) . \tag{8}$$

Summing these two inequalities leads to the resulting regret upper bound. The second bound comes since $\sum_{i=1}^{n} \sqrt{\mathbb{E}[x_{t,i,j}]} \le \sqrt{n}$. □

For the stochastic environment, it is key to prove a *self-bounding* constraint like previous works (Zimmert, Luo, and Wei 2019; Zimmert and Seldin 2019; Wei and Luo 2018). The common suboptimality gap of putting item i at position j is defined as $\Delta_{i,j} = \beta_j(\alpha_j - \alpha_i)$, the reward difference from the right item (Komiyama, Honda, and Takeda 2017). This could be negative for $i < j$. When this happens, a better item is put at position j. Then there must be some *bad* item placed before position j. We account for this situation and introduce a new suboptimality gap definition that is more suitable to PBM.

Definition 1. *For any* $i \in [n]$ *and* $j \in [m]$*, define*

$$\Delta_{i,j} = \begin{cases} (\beta_j - \beta_{j+1})(\alpha_j - \alpha_i) & j < i \,, \\ 0 & j = i \,, \\ (\beta_{j-1} - \beta_j)(\alpha_i - \alpha_j) & j > i \,. \end{cases}$$

The key idea for this definition comes from the incurred minimal regret of misplacing items. For the case that item i is put at position j with $i < j$. The item i is misplaced but is better than the right item at position j, which should be item j. This means there must be some *bad* items misplaced at earlier positions. The optimistic case is that item j is just put at one position ahead $j - 1$. Switching item i and item j would give the minimal regret gap, i.e. $\Delta_{i,j}$ is defined as the difference between the reward of $(\cdots, i, j, \cdots)$ and

$(\cdots, j, i, \cdots)$ where the only effectively involved positions are $j-1, j$. The case of $i > j$ is similar.

Now we can present the *self-bounding* constraint for PBM based on this introduced suboptimality gap.

Lemma 1. *For the stochastic environment, the regret satisfies*

$$R(T) \geq \frac{1}{2} \sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{j=1}^{m} \Delta_{i,j} \mathbb{E}[x_{t,i,j}] \,.$$

We first present Lemma 2 which reveals the property of the introduced suboptimality gap. The proof of Lemma 2 is postponed to Appendix.

Lemma 2. *Let $i_1, i_2, \ldots, i_m$ be any sequence chosen from $[n]$ without repetition. Then*

$$\sum_{j=1}^{m} (\alpha_j \beta_j - \alpha_{i_j} \beta_j) \geq \frac{1}{2} \sum_{j=1}^{m} \Delta_{i_j, j} \,.$$

Then Lemma 1 follows immediately from Lemma 2 by summing over the time horizon.

Proof of Lemma 1. Lemma 2 implies that

$$\begin{aligned} R(T) &= \sum_{t=1}^{T} \sum_{j=1}^{m} \mathbb{E}[\beta_j \alpha_j - \beta_j \alpha_{I_{t,j}}] \geq \sum_{t=1}^{T} \sum_{j=1}^{m} \mathbb{E}[\frac{1}{2} \Delta_{I_{t,j}, j}] \\ &= \frac{1}{2} \sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{j=1}^{m} \Delta_{i,j} \mathbb{E}[x_{t,i,j}] \,, \end{aligned}$$

which completes the proof. □

With the self-bounding constraint in Lemma 1, we can obtain the following regret bound for the stochastic setting.

Theorem 2. *For the stochastic environment, the regret of* FTRL-PBM *algorithm is upper bounded by*

$$\begin{aligned} R(T) &\leq \left(18 \sum_{j=1}^{m} \sum_{\substack{i=1 \\ i \neq j}}^{n} \frac{1}{\Delta_{i,j}} + 4m \right) \log T + 6m \\ &= O\Big(\frac{mn}{\Delta_\beta \Delta} \log(T) \Big) \,. \end{aligned}$$

This regret bound improves a factor of $O(m)$ over that of BOBW semi-bandits (Zimmert, Luo, and Wei 2019) which is designed for general combinatorial cases. Our regret upper bound is $O(\frac{1}{\Delta_\beta})$ worse than (Lattimore et al. 2018) which studies only the stochastic environments.

Proof.

$$\begin{aligned} R(T) \leq & 2R(T) - \frac{1}{2} \sum_{t=1}^{T} \sum_{j=1}^{m} \sum_{\substack{i=1 \\ i \neq j}}^{n} \mathbb{E}[x_{t,i,j}] \Delta_{i,j} \\ \leq & \sum_{t=1}^{T} \sum_{j=1}^{m} \sum_{\substack{i=1 \\ i \neq j}}^{n} \left(6 \sqrt{\frac{\mathbb{E}[x_{t,i,j}]}{t}} - \frac{1}{2} \mathbb{E}[x_{t,i,j}] \Delta_{i,j} \right) \\ & + 6m + 4m \log T \\ \leq & \sum_{t=1}^{T} \sum_{j=1}^{m} \sum_{\substack{i=1 \\ i \neq j}}^{n} \frac{18}{\Delta_{i,j} t} + 4m \log T + 6m \\ \leq & \left(18 \sum_{j=1}^{m} \sum_{\substack{i=1 \\ i \neq j}}^{n} \frac{1}{\Delta_{i,j}} + 4m \right) \log T + 6m \,, \end{aligned}$$

where the first inequality is by Lemma 1, the second inequality is by Eq.(7) and Eq.(8), and the third inequality is due to the AM–GM inequality. □

Lower Bound

We provide an improved lower bound for PBM and defer its proof to Appendix.

Theorem 3. *Suppose that $n \geq \max\{m+3, 2m\}$ and $T \geq n$. For any algorithm there exists an instance of OLTR under PBM such that*

$$R(T) \geq \frac{1}{16} m \sqrt{(n-m+1)T} \,.$$

Our lower bound improves $O(\sqrt{m})$ over the state-of-the-art lower bound (Lattimore et al. 2018) and matches our upper regret bound.

Experiments

This section compares the empirical performances of our FTRL-PBM with related baselines where TopRank (Lattimore et al. 2018), PBM-PIE (Lagrée, Vernade, and Cappé 2016), PMED (Komiyama, Honda, and Takeda 2017) are designed for the stochastic environment and RankedExp3 (Radlinski, Kleinberg, and Joachims 2008), MW (Kale, Reyzin, and Schapire 2010) are designed for the adversarial environment. We do not include (Zimmert, Luo, and Wei 2019) since we could not find an efficient method for PBM with their hybrid regularizer. Since the vanilla PBM-PIE in (Lagrée, Vernade, and Cappé 2016) needs the knowledge of the position examination probabilities, we use a bi-convex optimization to estimate the examination probabilities for PBM-PIE like (Komiyama, Honda, and Takeda 2017) rather than directly supplying. All parameters are kept the same as in their original papers. For all experiments, we use $n = 10$ items and $m = 5$ positions.

We only present the results of experiments on synthetic data in this section, The results of experiments on real-world data are deferred to Appendix. For the synthetic data, we set the position examination probabilities to $\beta = (1, \frac{1}{2}, \cdots, \frac{1}{5})$ which are commonly adopted in previous works (Wang et al. 2018; Li, Lattimore, and Szepesvári 2019). The attractiveness of items are set as $\alpha = (0.95, 0.95 - \Delta, 0.95 - 2\Delta, \cdots, 0.95 - 9\Delta)$. We consider two cases of $\Delta = 0.03$ and $\Delta = 0.01$.

We first construct stochastic environments (Lagrée, Vernade, and Cappé 2016; Komiyama, Honda, and Takeda 2017) from the item attractiveness and position examination probabilities set above. The results are shown in Fig.1(a)(d).

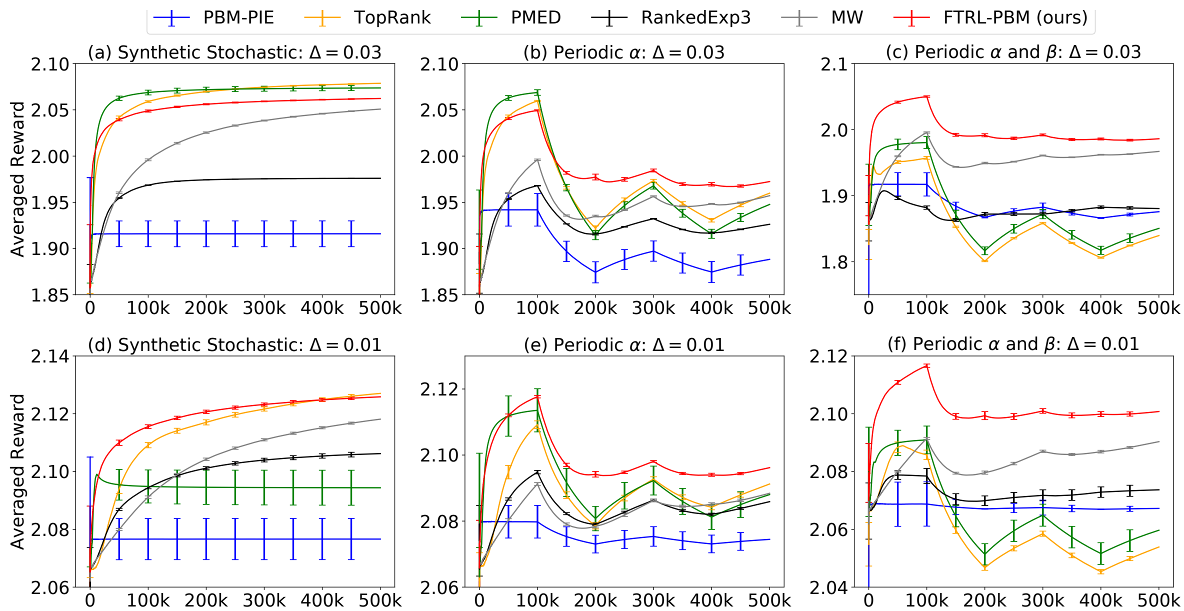

Figure 1: This figure shows empirical comparisons between our FTRL-PBM and `TopRank`, `PBM-PIE`, `PMED`, `RankedExp3` and `MW` in stochastic and periodic environments constructed by synthetic data. We adopt the metric of averaged rewards, which are the cumulative rewards divided by time t. All results are averaged over 10 random runs and the error bars are standard errors which are standard deviations divided by $\sqrt{10}$.

Next we construct adversarial environments. Since it is a bit hard to design a real adversarial environment, we construct two periodical stochastic environments like (Zimmert and Seldin 2019; Zimmert, Luo, and Wei 2019)[1]. We divide the whole time horizon into phases with 100k rounds in each phase. For the first periodic environment, the position examination probabilities are fixed and the attractiveness of the first five items and last five items are exchanged periodically. Specifically, the odd phase uses the same environment as in the stochastic case and the even phase uses item attractiveness $(0.95-5\Delta,\cdots,0.95-9\Delta,0.95,\cdots,0.95-4\Delta)$. For the second periodic environment, the position examination probabilities are also changed periodically. We reverse the order of α,β simultaneously. Specifically, the odd phase uses the same environment as in the stochastic case and the even phase uses item attractiveness $(0.95-9\Delta,0.95-8\Delta,\cdots,0.95-\Delta,0.95)$ and position examination probabilities $(\frac{1}{5},\frac{1}{4},\cdots,1)$. The results are shown in Fig.1(b)(c)(e)(f).

`TopRank` performs best in (a)(d) since it is specially designed for the stochastic environment. `PMED` is also designed for the stochastic environment and has almost the same performance as `TopRank` in (a). Since it needs to solve a bi-convex optimization problem, fixed iterations would not give good convergence. Thus its performance has a large variance and deteriorates a lot for small gap (d). The design of `PBM-PIE` needs the knowledge of position bias. Though we can estimate them by bi-convex optimization, the estimation error would be amplified in the results when the estimated values are adopted directly. Then the performance of `PBM-PIE` is not very good and has a large variance in (a)(d). All of `TopRank`, `PMED` and `PBM-PIE` are strongly affected by the periodic changes (b)(c)(e)(f).

Our algorithm FTRL-PBM is competitive in stochastic environments and is best in adversarial environments, showing the robustness and simultaneous learning ability of our algorithm. The greedy idea in `RankedExp3` is not very suitable for PBM but since it is designed for the adversarial environment, its performances are relatively stable. `MW` can be regarded as FTRL with negative Shannon entropy. It has good performances in some scenarios but is not the best due to the suboptimality of the regularizer.

[1]They assume the relative order of items is fixed and the phase length is increasing.

Conclusions

To characterize the dynamic changes of online learning to rank (OLTR) environments, we study how to simultaneously learn in both stochastic and adversarial environments for OLTR under the position-based model (PBM). We design an algorithm based on the follow-the-regularized-leader framework and prove its efficiency in both environments. We also provide a lower bound for adversarial PBM which matches our upper bound. Experiments also validate the robustness of our algorithm.

Our results only focus on PBM. It would be a practical and promising topic to design efficient algorithms for both environments under general click models subsuming multiple click models. Further, the adversarial setting for general click models is open and suggested to be solved in the future.

Acknowledgments

The corresponding author Shuai Li is supported by National Natural Science Foundation of China (62006151, 62076161). This work is sponsored by Shanghai Sailing Program. Cheng Chen is supported by Singapore Ministry of Education (AcRF) Tier 2 grant MOE2018-T2-1-013.

References

Ailon, N. 2014. Improved bounds for online learning over the permutahedron and other ranking polytopes. In *Artificial Intelligence and Statistics*, 29–37. PMLR.

Ailon, N.; Hatano, K.; and Takimoto, E. 2016. Bandit online optimization over the permutahedron. *Theoretical Computer Science*, 650: 92–108.

Audibert, J.-Y.; Bubeck, S.; and Lugosi, G. 2014. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1): 31–45.

Auer, P.; and Chiang, C.-K. 2016. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*, 116–120. PMLR.

Bauschke, H. H.; and Lewis, A. S. 2000. Dykstras algorithm with bregman projections: A convergence proof. *Optimization*, 48(4): 409–427.

Besson, L.; and Kaufmann, E. 2018. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*.

Bregman, L. 1967. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *Ussr Computational Mathematics and Mathematical Physics*, 7(3): 200–217.

Bubeck, S.; and Cesa-Bianchi, N. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations & Trends in Machine Learning*.

Bubeck, S.; and Slivkins, A. 2012. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, 42–1. JMLR Workshop and Conference Proceedings.

Chen, W.; Wang, Y.; and Yuan, Y. 2013. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, 151–159. PMLR.

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

Cohen, A.; and Hazan, T. 2015. Following the perturbed leader for online structured learning. In *International Conference on Machine Learning*, 1034–1042. PMLR.

Combes, R.; Magureanu, S.; and Proutiere, A. 2017. Minimal exploration in structured stochastic bandits. In *Conference on Neural Information Processing Systems*, 1761–1769.

Combes, R.; Talebi, S.; Proutière, A.; and Lelarge, M. 2015. Combinatorial Bandits Revisited. In *Conference on Neural Information Processing Systems*.

Cutkosky, A.; and Boahen, K. A. 2017. Stochastic and Adversarial Online Learning without Hyperparameters. In *NIPS*, 5059–5067.

Dunn, J. C.; and Harshbarger, S. 1978. Conditional gradient algorithms with open loop step size rules. *Journal of Mathematical Analysis and Applications*, 62(2): 432–444.

Gai, Y.; Krishnamachari, B.; and Jain, R. 2012. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, 20(5): 1466–1478.

Helmbold, D. P.; and K Warmuth, M. 2009. Learning Permutations with Exponential Weights. *Journal of Machine Learning Research*, 10(7).

Jin, T.; and Luo, H. 2020. Simultaneously Learning Stochastic and Adversarial Episodic MDPs with Known Transition. *Advances in Neural Information Processing Systems*, 33.

Kale, S.; Reyzin, L.; and Schapire, R. E. 2010. Non-stochastic bandit slate problems. In *Advances in Neural Information Processing Systems*, 1054–1062.

Katariya, S.; Kveton, B.; Szepesvari, C.; Vernade, C.; and Wen, Z. 2017. Stochastic rank-1 bandits. In *Artificial Intelligence and Statistics*, 392–401. PMLR.

Komiyama, J.; Honda, J.; and Takeda, A. 2017. Position-based multiple-play bandit problem with unknown position bias. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 5005–5015.

Koolen, W. M.; Grünwald, P.; and van Erven, T. 2016. Combining adversarial guarantees and stochastic fast rates in online learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 4464–4472.

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

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

Lagrée, P.; Vernade, C.; and Cappé, O. 2016. Multiple-play bandits in the position-based model. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 1605–1613.

Lattimore, T.; Kveton, B.; Li, S.; and Szepesvári, C. 2018. TopRank: a practical algorithm for online stochastic ranking. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 3949–3958.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit algorithms*. Cambridge University Press.

Lee, C.-W.; Luo, H.; Wei, C.-Y.; Zhang, M.; and Zhang, X. 2021. Achieving Near Instance-Optimality and Minimax-Optimality in Stochastic and Adversarial Linear Bandits Simultaneously. *arXiv preprint arXiv:2102.05858*.

Li, S.; Lattimore, T.; and Szepesvári, C. 2019. Online learning to rank with features. In *International Conference on Machine Learning*, 3856–3865. PMLR.

Li, S.; Wang, B.; Zhang, S.; and Chen, W. 2016. Contextual combinatorial cascading bandits. In *International conference on machine learning*, 1245–1253. PMLR.

Liu, T.-Y.; et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3): 225–331.

Mendelsohn, N. S.; and Dulmage, A. L. 1958. The convex hull of sub-permutation matrices. *Proceedings of the American Mathematical Society*, 9(2): 253–254.

Mourtada, J.; and Gaïffas, S. 2019. On the optimality of the Hedge algorithm in the stochastic regime. *Journal of Machine Learning Research*, 20: 1–28.

Neu, G. 2015. First-order regret bounds for combinatorial semi-bandits. In *Conference on Learning Theory*, 1360–1375. PMLR.

Neu, G.; and Bartók, G. 2013. An efficient algorithm for learning with semi-bandit feedback. In *International Conference on Algorithmic Learning Theory*, 234–248. Springer.

Radlinski, F.; Kleinberg, R.; and Joachims, T. 2008. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th international conference on Machine learning*, 784–791.

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

Seldin, Y.; and Lugosi, G. 2017. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory*, 1743–1759. PMLR.

Seldin, Y.; and Slivkins, A. 2014. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, 1287–1295. PMLR.

Slivkins, A.; Radlinski, F.; and Gollapudi, S. 2013. Ranked bandits in metric spaces: learning diverse rankings over large document collections. *Journal of Machine Learning Research*, 14(Feb): 399–436.

Wang, X.; Golbandi, N.; Bendersky, M.; Metzler, D.; and Najork, M. 2018. Position bias estimation for unbiased learning to rank in personal search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 610–618.

Wei, C.-Y.; and Luo, H. 2018. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, 1263–1291. PMLR.

Zimmert, J.; Luo, H.; and Wei, C.-Y. 2019. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *International Conference on Machine Learning*, 7683–7692. PMLR.

Zimmert, J.; and Seldin, Y. 2019. An optimal algorithm for stochastic and adversarial bandits. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 467–475. PMLR.

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

Zong, S.; Ni, H.; Sung, K.; Ke, N. R.; Wen, Z.; and Kveton, B. 2016. Cascading bandits for large-scale recommendation problems. *arXiv preprint arXiv:1603.05359*.