Dynamic Ensemble of Contextual Bandits to Satisfy Users' Changing Interests

Qingyun Wu
University of Virginia
Charlottesville, VA , USA
qw2ky@virginia.edu

Huazheng Wang
University of Virginia
Charlottesville, VA , USA
hw7ww@virginia.edu

Yanen Li
Snap Inc.
Los Angeles, CA, USA
yanen.li@snap.com

Hongning Wang
University of Virginia
Charlottesville, VA, USA
hw5x@virginia.edu

ABSTRACT

Recommender systems have to handle a highly non-stationary environment, due to users' fast changing interests over time. Traditional solutions have to periodically rebuild their models, despite high computational cost. But this still cannot empower them to automatically adjust to abrupt changes in trends caused by timely information. It is important to note that the changes of reward distributions caused by a non-stationary environment can also be *context dependent*. When the change is orthogonal to the given context, previously maintained models should be reused for better recommendation prediction.

In this work, we focus on contextual bandit algorithms for making adaptive recommendations. We capitalize on the unique context-dependent property of reward changes to conquer the challenging non-stationary environment for model update. In particular, we maintain a dynamic ensemble of contextual bandit models, where each bandit model's reward estimation quality is monitored regarding given context and possible environment changes. Only the admissible models to the current environment will be used for recommendation. We provide a rigorous upper regret bound analysis of our proposed algorithm. Extensive empirical evaluations on both synthetic and three real-world datasets confirmed the algorithm's advantage against existing non-stationary solutions that simply create new models whenever an environment change is detected.

CCS CONCEPTS

• **Information systems → Recommender systems**; • **Theory of computation → Sequential decision making**; **Online learning algorithms**; **Regret bounds**;

KEYWORDS

Non-stationary bandits; recommender systems; regret analysis

ACM Reference Format:
Qingyun Wu, Huazheng Wang, Yanen Li, and Hongning Wang. 2019. Dynamic Ensemble of Contextual Bandits to Satisfy Users' Changing Interests. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3308558.3313727

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313727

1 INTRODUCTION

The overwhelming volume of online content make personalized recommendation an indispensable component in modern information service systems. Traditional solutions, including content-based filtering [17, 26], collaborative filtering [6, 30], and hybrid approaches [22], provide recommendations by leveraging users' interests as demonstrated in their past activities. However, in many practical applications, such as news recommendation, both content popularity and user interests evolve frequently over time, which make the traditional offline learning approaches incompetent [18].

In recent years, multi-armed bandits, and more specifically contextual bandits, have become a referenced online learning solution to deal with this dynamic nature of recommendations [1, 20, 27, 32, 33, 36, 37]. Contextual bandit solutions explore the unknowns by collecting users' feedback in real time to estimate the utility/reward of new content with available side-information or context information. They provide a principled way to find optimal trade-offs between exploration and exploitation [3, 4], and have been successfully deployed in many important practical scenarios [20, 23].

However, most existing contextual bandit algorithms assume a stationary environment [4, 20], where the expected reward on each arm is drawn from an unknown yet fixed reward mapping function based on the given context. This assumption, however, rarely holds in many real-world applications, where the underlying reward mapping undergoes slow or abrupt changes due to various factors. For example, Liu et al. [24] reported that readers' preferences over news articles shift with time and events in Google News. By analyzing more than 20,000 twitter users over a four-months period, Abel et al. [2] found that the interests of individual users into a topic evolve differently over time. More importantly, the duration during which users are interested in an event-like topic differs significantly among each other. In other words, users' interest change over time but the change is unknown beforehand.

In this work, we focus on the setting where there are unknown (to the learner) and abrupt changes in terms of user preferences. Between consecutive changes, the reward distribution remains stationary yet unknown, i.e., *piecewise stationary*. Existing bandit algorithms address such a dynamic environment by either introducing a forgetting mechanism to downweight historical observations [12, 15], or creating a bandit model for each newly detected stationary period [7, 14, 35, 38]. These strategies, nevertheless, failed to recognize an important property of this non-stationary environment: the changes of user interests can be *context dependent*. Even though one's interest could change frequently, his/her preference on a particular type of items might be stable over a longer period of time. For example, in the news recommendation scenario mentioned above, users' preferences over sports news may change with

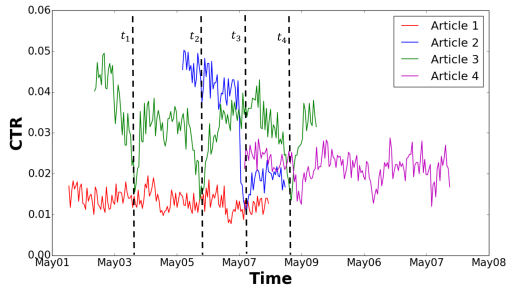


Figure 1: Real-time click-through-rate of four sample news articles collected from Yahoo user click log dataset [20, 21].

sport seasons. However, at the same time, their preferences over political news may stay stationary, independent of the sport season.

This phenomenon can be verified in Figure 1, where we collect real-time click-through-rate (CTR) of four sample news articles from the public Yahoo front-page user click log [20, 21] over a week's period. In the figure, each point denotes average CTR over 4000 log records. We can clearly observe the changes of user interest on article 2 at time t_3 and on article 3 at time t_1, t_2 and t_4 in this period. In the meanwhile, the CTR of article 1 and 4 remain quite stable, i.e., the expected reward for recommending these two articles remain unchanged. As a result, the experience recorded in old models can still be used to make accurate reward estimations for these two articles. On the contrary, strategies that discount observations or abandon the "old" models must regain confidence in their newly estimated parameters, which many lead to higher regret due to redundant explorations. The key is thus to recognize the reward change in a per-arm basis regarding the context.

To capitalize on such a unique property of the changing environment, we develop our contextual bandit algorithm that adapts its arm selection and model update strategy with regard to users' interest changes in a context-dependent manner. Our solution consists of a dynamic set of contextual bandit models, collectively referred to as *bandit experts*, which are maintained to estimate the underlying reward distribution. In the meanwhile, we monitor each bandit expert's reward estimation quality regarding specific context through another bandit model, referred to as *bandit auditors*. The auditor predicts whether a bandit expert is 'admissible' to a specific arm under the given context. At each round of interaction, an ensemble of admissible bandit experts is created to estimate the reward of each arm; and the arm with the highest upper confidence bound of estimated reward will be selected. The acquired feedback is used to update all admissible bandit experts for this chosen arm, and their corresponding auditors. When no admissible bandit expert exists, a new bandit expert will be created and added to the set before evaluating the recommendation candidates.

We rigorously prove a sublinear upper regret bound of our proposed algorithm, which guarantees that the number of sub-optimal recommendations from our algorithm reduces rapidly over time. We also show that if one fails to model the context dependent changes, a much worse upper regret bound is inevitable. Extensive empirical evaluations in one synthetic and three real-world datasets also confirmed the effectiveness of the proposed algorithm, especially in handling the changing popularity and user preferences in practice.

2 RELATED WORK

In this work, we focus on the stochastic setting of multi-armed bandit problems [29]. Under this setting, various exploration strategies have been proposed to conquer the exploit-explore dilemma, such as upper confidence bound [3, 4, 20], epoch greedy [19] and posterior sampling [9, 16]. However, most of these algorithms assume a static environment, which is often violated in practice. A number of bandit models hence have been proposed to deal with non-stationary environments, and there are mainly two types of environment changes studied: gradual changes and abrupt changes.

For gradual changes, Whittle [34] introduced the restless bandits, where the states of arms can change in each step according to a stochastic transition function. But this setting is notoriously intractable [5, 25, 31].

Our work concerns abrupt changes of environment, where most solutions only focus on context-free bandits. Hartland et al. [15], and Garivier and Moulines [12] proposed a γ-Restart algorithm and a discounted-UCB algorithm, in which a discount factor is introduced to exponentially decay the effect of past observations. Both algorithms need properly designed hyper-parameters, such as the discount factor. Yu and Mannor [38] proposed a windowed mean-shift detection algorithm to detect the changes in the environment. They assume that at each iteration, the agent can not only get feedback from the selected arms but also query a subset of arms for additional observations, which would be very expensive in practice. Raj and Kalyani proposed discounted Thompson Sampling in [28]. Recently Cao et al. [7] proposed a UCB based algorithm with change detection module to detect changes and restart exploration accordingly. Liu et al. [10] proposed to use cumulative sum and Page-Hinkley test to detect sudden changes in the environment. A sublinear upper regret bound is proved for a simplified Bernoulli bandit environment with a strong detectability assumption. All of the aforementioned solutions are non-contextual bandit solutions, which are incapable to leverage the rich side information in real-world applications.

Some recent works have realized the lack of contextual bandits in a non-stationary environment, and extended corresponding solutions. Hariri et al. [14] proposed a contextual Thompson sampling algorithm with a change detection module, which involves iteratively applying a combination of cumulative sum charts and bootstrapping to capture potential changes. But due to its empirical nature, no theoretical property is known about it. Wu et al. [35] developed a two-level hierarchical bandit algorithm, which detects and adapts to changes in the environment by maintaining a suite of contextual bandit models. Regret analysis is provided under a strong assumption about the change. However, to the best of our knowledge, among the existing non-stationary bandit solutions, no work utilizes the context dependent property of reward changes in a piece-wise stationary environment. Hence, none of them is able to exploit the existence of *context dependent changes*. Our algorithm avoids the unnecessary penalties in regret incurred by indiscriminately dismissing existing bandit models and tries to reuse the existing models when judged 'admissible.'

3 METHODOLOGY

In this section, we first introduce our notations and assumptions about the non-stationary environment, then illustrate our proposed algorithm, followed with a rigorous regret bound analysis.

3.1 Problem Setups

In a multi-armed bandit problem, a learner sequentially selects an arm a_t from a candidate pool $\mathcal{A} = \{a_1, a_2, ..., a_k\}$ to interact with the environment, and receives the corresponding reward r_{a_t}. The goal of the learner is to maximize its accumulated reward over a finite time horizon T. In a stochastic contextual bandit setting, each candidate arm a is associated with a feature vector $\mathbf{x}_a$ (assume $\|\mathbf{x}_a\|_2 \leq 1$ without loss of generality), referred as context. The corresponding reward r_a is determined by the context vector and an underlying bandit parameter $\boldsymbol{\theta}^*$ (assume $\|\boldsymbol{\theta}^*\|_2 \leq 1$ without loss of generality). When applying to the recommendation scenario, the bandit parameter $\boldsymbol{\theta}^*$ can be interpreted as the underlying parameter that controls users' interests and $\mathbf{x}_a$ is the available context information about the candidate item a. Most existing contextual bandit algorithms consider $\boldsymbol{\theta}^*$ as constant over time [1, 20], which imposes a strong stationary assumption about the environment.

In this work, firstly, we relax this stationary assumption by allowing abrupt changes in $\boldsymbol{\theta}^*$: the ground-truth $\boldsymbol{\theta}^*$ changes arbitrarily at unknown time points, but remains constant between any two consecutive change points [12, 14, 15] as follows:

$$\underbrace{r_0, r_1, \cdots, r_{t_{c_1}-1}}_{\text{governed by } \boldsymbol{\theta}^*_{c_0}}, \underbrace{r_{t_{c_1}}, r_{t_{c_1}+1}, \cdots, r_{t_{c_2}-1}}_{\text{governed by } \boldsymbol{\theta}^*_{c_1}}, \cdots, \underbrace{r_{t_{c_{\Gamma_T}}}, r_{t_{c_{\Gamma_T}}+1}, \cdots, r_T}_{\text{governed by } \boldsymbol{\theta}^*_{c_{\Gamma_T}}}$$

where the change points $\{t_c\}_{c=c_1}^{c_{\Gamma_T}}$ and the corresponding bandit parameters $\{\boldsymbol{\theta}^*_c\}_{c=c_0}^{c_{\Gamma_T}}$ are unknown to the learner. We only assume there are at most Γ_T change points in the environment up to time T, with $\Gamma_T \ll T$. To simplify the discussion of our developed algorithm, a linear reward structure is postulated, but it can be readily extended to more complicated structures, such as generalized linear models [11]. Specifically, we have $r_{a,t} = \mathbf{x}_a^\top \boldsymbol{\theta}^*_t + \eta_t$, where η_t is Gaussian noise drawn from $N(0, \sigma^2)$.

Secondly, we capitalize on the unique property that in a contextual bandit setting the changes of reward distribution are context-dependent. Thus we categorize arms into two types, *change-invariant* and *change-sensitive*, between any two stationary periods. For stationary periods i and j with their ground-truth bandit parameters $\boldsymbol{\theta}^*_i$ and $\boldsymbol{\theta}^*_j$, the arm that satisfies $|\mathbf{x}_a^\top \boldsymbol{\theta}^*_i - \mathbf{x}_a^\top \boldsymbol{\theta}^*_j| \leq \Delta_L$ (with $\Delta_L > 0$) is referred as a change-invariant arm; otherwise as a change-sensitive arm. Δ_L is a parameter to introduce flexibility and accommodate stochastic noise, which relaxes the requirement that the change has to be completely orthogonal to the context vector of a change-invariant arm. This makes our problem setting more general than those in [10, 38]. An illustration of these two types of arms is provided in Figure 2: when the ground-truth bandit parameter changes from $\boldsymbol{\theta}^*_c$ to $\boldsymbol{\theta}^*_{c+1}$, although there are *change-sensitive arms*, for example arm a_1, whose expected reward changes dramatically, there are also *change-invariant arms*, for example arm a_2, whose context vectors are orthogonal to the change, i.e., $\mathbf{x}_a^\top(\boldsymbol{\theta}^*_c - \boldsymbol{\theta}^*_{c+1}) = 0 < \Delta_L$, so that their expected reward does not change significantly even after the environment changes. For example, mapping it back to our previous news recommendation example, if the reward change in a user was caused by the change of sports season, the sports news and political news could be considered as the change-sensitive and change-invariant arms respectively for this user.

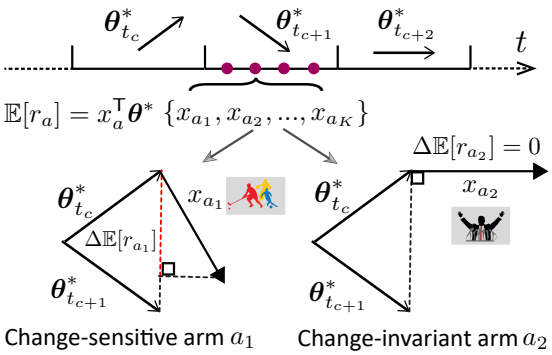


Figure 2: An illustrative example of context-dependent changes in a piecewise stationary environment. A linear reward assumption is postulated to simplify the illustration.

To differentiate the actual reward change from stochastic noise, we impose the following assumption about the non-stationary environment, which characterizes the detectability of reward changes between the stationary period i and j,

Assumption 1. *Among the change-sensitive arms between stationary period i and j, there are at least ρ portion of them satisfying $|\mathbf{x}_a^\top \boldsymbol{\theta}^*_i - \mathbf{x}_a^\top \boldsymbol{\theta}^*_j| > \Delta_H$.*

This assumption requires that between any two stationary periods, there are a number of change-sensitive arms undergo perceivable reward changes, which differentiate these two periods.

3.2 A Dynamic Ensemble of Bandits

In the non-stationary environment specified above, where the changes of users' interests occasionally happen at unknown time points, new bandit models should be rebuilt in accordance with the changes of underlying user interest. In addition, the possible existence of change-invariant arms urges us to reuse the bandit models estimated for those earlier periods, so that more accurate reward estimation on such arms can be achieved sooner so as to obtain reduced regret in a new stationary period.

In order to achieve these goals, three challenges have to be addressed: 1) as the changes of environment are unknown to the learner, how to detect the potential change of user interests and create new bandit models to account for the change-sensitive arms in a new environment? 2) as the reward changes in each arm are context-dependent, how to recognize the change-invariant arms at current period such that experience from old models can be fully utilized? and 3) which arm to choose given multiple bandit models might exist at the same time? To avoid any potential ambiguity in our later discussions, we refer to the contextual bandit models created for reward estimation as *bandit experts*.

We address the first two challenges by creating a companion bandit model for each bandit expert to monitor its reward estimation quality. We refer to this companion bandit model as a *bandit auditor*. In a nutshell, a bandit expert, who works with the context features and reward collected from its chosen arms, is responsible

for reward estimation regarding its corresponding environment. Its companion bandit auditor works with the context features of the expert's choices and the observed prediction errors from the bandit expert. The auditor is responsible for assessing the expert's prediction accuracy. At each round, every bandit auditor evaluates whether the monitored bandit expert is *admissible* to make an accurate reward estimation for a given arm, with respect to potential environment changes. A bandit expert being identified as admissible to a particular arm indicates with a high probability that, either 1) no change has happened since the creation of this bandit expert, or 2) the environment has changed but the arm is change-invariant between this bandit expert's estimated reward distribution and the current period's underlying reward distribution. This addresses the second challenge. When no admissible bandit expert exists for a given arm, it is thus highly likely to be a change-sensitive arm in a new environment, and a new bandit expert is needed. This addresses the first challenge.

To address the third challenge, at each round of interaction, following the principle of optimism in the face of uncertainty [1, 20], an arm is chosen by the upper confidence bound of reward estimation based on an ensemble of all its admissible bandit experts. The acquired feedback for the selected arm/item is used to update all corresponding bandit experts and their auditors. We name the resulting bandit algorithm as Dynamic Ensemble of Bandit Experts, or DenBand in short. We describe DenBand in Algorithm 1 [1] and discuss the key components of it in details as follows.

Bandit Expert. Define t_m as the time when bandit expert m is created. Each bandit expert m maintains an estimated bandit parameter $\hat{\boldsymbol{\theta}}_t(m)$ for the stationary period at t_m. Define $\mathcal{I}_t^{\theta}(m)$ as a set of timestamps when observation $(\mathbf{x}_{a_i}, r_{a_i,i})$ is assigned to the bandit expert m for model update till time t (in line 24-25 of Algorithm 1). $r_{a_i,i}$ is the observed reward on arm a_i at time i. Because of our linear reward structure, $\hat{\boldsymbol{\theta}}_t(m)$ can be readily estimated by $\hat{\boldsymbol{\theta}}_t(m) = \mathbf{A}_t^{-1}(m)\mathbf{b}_t(m)$, in which $\mathbf{A}_t(m) = \lambda\mathbf{I} + \sum_{i\in\mathcal{I}_t^{\theta}(m)} \mathbf{x}_{a_i}\mathbf{x}_{a_i}^{\mathsf{T}}$, $\mathbf{I}$ is a $d \times d$ identity matrix, λ is the regularization coefficient in the least square regression, and $\mathbf{b}_t(m) = \sum_{i\in\mathcal{I}_t^{\theta}(m)} \mathbf{x}_{a_i} r_{a_i,i}$.

Bandit Auditor. Denote the reward estimation error of bandit expert m on arm a at time t as $e_{a,t}(m) = \hat{r}_{a,t}(m) - r_{a,t}$, in which $\hat{r}_{a,t}(m) = \mathbf{x}_a^{\mathsf{T}}\hat{\boldsymbol{\theta}}_t(m)$. We have $\mathbb{E}[e_{a,t}(m)] = \mathbf{x}_a^{\mathsf{T}}(\hat{\boldsymbol{\theta}}_t(m) - \boldsymbol{\theta}_t^*)$, which is referred as 'badness' of bandit expert m on arm a at time t and leads to a linear structure for badness estimation. We create a new bandit model with the target parameter for estimation as $\boldsymbol{\beta}_t^*(m) = \hat{\boldsymbol{\theta}}_t(m) - \boldsymbol{\theta}_t^*$, and refer to it as the bandit auditor of bandit expert m. We maintain and update the bandit auditors in a similar manner as that in bandit experts. Denote $\mathcal{I}_t^{\beta}(m)$ as a set of timestamps when observation $\left(\mathbf{x}_{a_i}, e_{a_i,i}(m)\right)$ is assigned to the bandit auditor for bandit expert m up to time t (line 21-23 in Algorithm 1). The bandit auditor estimates $\boldsymbol{\beta}_t^*(m)$ by $\hat{\boldsymbol{\beta}}_t(m) = \mathbf{C}_t^{-1}(m)\mathbf{d}_t(m)$, in which $\mathbf{C}_t(m) = \lambda\mathbf{I} + \sum_{i\in\mathcal{I}_t^{\beta}(m)} \mathbf{x}_{a_i}\mathbf{x}_{a_i}^{\mathsf{T}}$ and $\mathbf{d}_t(m) = \sum_{i\in\mathcal{I}_t^{\beta}(m)} \mathbf{x}_{a_i} e_{a_i,i}(m)$. Intuitively, the bandit auditor for expert m evaluates whether an arm a at time t is change-invariant to the reward distributions specified by $\boldsymbol{\theta}_t^*$ and $\boldsymbol{\theta}_{t_m}^*$. The definition of badness requires us to update

[1] Open source implementation of DenBand can be found in https://github.com/huazhengwang/BanditLib

Algorithm 1 Dynamic Ensemble of Bandit Experts (DenBand)

1: **Inputs:** $\alpha \in \mathbb{R}_+$, $\lambda > 0$, δ_1, $\delta_2 \in (0, 1)$, τ, Δ_L
2: **Initialize:** Create and initialize bandit expert m: $\mathbf{A}_1(m) = \lambda\mathbf{I}$, $\mathbf{b}_1(m) = \mathbf{0}$, $\hat{\boldsymbol{\theta}}_1(m) = \mathbf{0}$, and its auditor: $\mathbf{C}_1(m) = \lambda\mathbf{I}$, $\mathbf{d}_1(m) = \mathbf{0}$, $\hat{\boldsymbol{\beta}}_1(m) = \mathbf{0}$. Initialize the bandit expert set $\mathcal{M}_1 = \{m\}$
3: **for** $t = 1$ to T **do**
4: **for** $a \in \mathcal{A}_t$ **do**
5: Create an admissible model set for arm a: $\mathcal{M}_t^a = \emptyset$
6: **for** $m \in \mathcal{M}_t$ **do**
7: $\hat{e}_{a,t}(m) = \mathbf{x}_a^{\mathsf{T}}\hat{\boldsymbol{\beta}}_t(m)$
8: Compute $B_{a,t}^{\beta}(m)$ and $B_{a,t}^{\theta}(m)$ by Eq (1) and (2)
9: **if** $|\hat{e}_{a,t}(m)| < B_{a,t}^{\theta}(m) + B_{a,t}^{\beta}(m) + \Delta_L$ **then**
10: Add m into $\mathcal{M}_t^a$
11: **end if**
12: **end for**
13: **if** $|\mathcal{M}_t^a| = 0$ **then**
14: Create and initialize a new bandit expert m and its auditor as in Line 2; Add m to $\mathcal{M}_t^a$ and $\mathcal{M}_t$
15: **end if**
16: Compute $\mathrm{UCB}_{a,t}$ of arm a with bandit experts in $\mathcal{M}_t^a$
17: **end for**
18: Select an arm by $a_t = \arg\max_{a\in\mathcal{A}} \mathrm{UCB}_{a,t}$
19: Observe reward $r_{a_t,t}$
20: **for** $m \in \mathcal{M}_t^{a_t}$ **do**
21: $\hat{r}_{a_t,t}(m) = \mathbf{x}_{a_t}^{\mathsf{T}}\hat{\boldsymbol{\theta}}_t(m)$, $e_{a_t,t}(m) = \hat{r}_{a_t,t}(m) - r_{a_t,t}$
22: Update $\{e_{a_i,i}\}_{i\in\mathcal{I}_t^{\beta}(m)}$ according to $\hat{\boldsymbol{\theta}}_t(m)$, and keep the size of $\mathcal{I}_t^{\beta}(m)$ to τ
23: $\mathbf{C}_{t+1}(m) = \mathbf{C}_t(m) + \mathbf{x}_{a_t}\mathbf{x}_{a_t}^{\mathsf{T}}$, $\mathbf{d}_{t+1}(m) = \sum_{i\in\mathcal{I}_t^{\beta}(m)} \mathbf{x}_{a_i} e_{a_i,i}$
24: **if** $m \in \mathcal{M}_t^{a_t}$ and $|\hat{r}_{a_t,t}(m) - r_{a_t,t}| \le B_{a_t,t}^{\theta}(m) + \Delta_L + \epsilon$ **then**
25: $\mathbf{A}_{t+1}(m) = \mathbf{A}_t(m) + \mathbf{x}_{a_t}\mathbf{x}_{a_t}^{\mathsf{T}}$, $\mathbf{b}_{t+1}(m) = \mathbf{b}_t(m) + \mathbf{x}_{a_t} r_{a_t,t}$
26: **else**
27: $\mathbf{A}_{t+1}(m) = \mathbf{A}_t(m)$, $\mathbf{b}_{t+1}(m) = \mathbf{b}_t(m)$
28: **end if**
29: $\hat{\boldsymbol{\theta}}_{t+1}(m) = \mathbf{A}_{t+1}(m)^{-1}\mathbf{b}_{t+1}(m)$
30: $\hat{\boldsymbol{\beta}}_{t+1}(m) = \mathbf{C}_{t+1}(m)^{-1}\mathbf{d}_{t+1}(m)$
31: **end for**
32: **end for**

$e_{a,t}(m)$ of all observations in $\mathcal{I}_t^{\beta}(m)$ whenever $\hat{\boldsymbol{\theta}}_t(m)$ is updated or $\boldsymbol{\theta}_t^*$ is changed. But as the environment change is unknown to the learner, this update is infeasible. We decide to only accumulate the most recent τ observations in $\mathcal{I}_t^{\beta}(m)$ for auditor update, and prove this still provides us a high probability bound of each bandit auditor's badness estimation in our regret analysis,

$$|\hat{e}_{a,t}(m) - \mathbb{E}[e_{a,t}(m)]| \le B_{a,t}^{\beta}(m) \tag{1}$$

where $B_{a,t}^{\beta}(m) = \left(\sigma^2\sqrt{d\ln(\frac{\lambda+|\mathcal{I}_t^{\beta}(m)|}{\lambda\delta_2})} + \sqrt{\lambda}\right)\|\mathbf{x}_a\|_{\mathbf{C}_t^{-1}(m)}$.

Bandit Expert Selection. Based on Eq (1) and our badness definition, we have $|\hat{e}_{a,t}(m)| \le \mathbb{E}[e_{a,t}(m)] + B_{a,t}^{\beta}(m) \le |\mathbf{x}_a^{\mathsf{T}}\hat{\boldsymbol{\theta}}_t(m) - \mathbf{x}_a^{\mathsf{T}}\boldsymbol{\theta}_{t_m}^*| + |\mathbf{x}_a^{\mathsf{T}}\boldsymbol{\theta}_{t_m}^* - \mathbf{x}_a^{\mathsf{T}}\boldsymbol{\theta}_t^*| + B_{a,t}^{\beta}(m)$. The first part on the right-hand side of this inequality is the reward estimation quality of bandit expert m, which can be bounded by,

$$|\mathbf{x}_a\hat{\boldsymbol{\theta}}_t(m) - \mathbf{x}_a\boldsymbol{\theta}_{t_m}^*| \le B_{a,t}^{\theta}(m,a) \tag{2}$$

where $B^{\theta}_{a,t}(m,a) = \left(3\sigma^2\sqrt{d\ln(\frac{\lambda+|\mathcal{I}^{\theta}_t(m)|}{\lambda\delta_1})} + \sqrt{\lambda}\right)\|\mathbf{x}_a\|_{\mathbf{A}^{-1}_t(m)}$. We should note that this bound is different from those for classical linear bandit algorithms (e.g., [1]), since it also has to account for the possible contamination from change-invariant arms (as we do not restrict Δ_L to zero). The second part on the right-hand side can be bounded by Δ_L, if no change has happened since t_m or the arm a is change-invariant between t_m and t. As a result, the condition $|\hat{e}_{a,t}(m)| < B^{\theta}_{a,t}(m) + B^{\beta}_{a,t}(m) + \Delta_L$ in line 9 of Algorithm 1 determines if the bandit expert m is admissible to arm a at time t (supported by Lemma 1 and 2 in Section 3.3). When there is no admissible bandit expert for arm a, a new bandit expert needs to be created to account for the detected change of environment (line 13-15 in Algorithm 1).

Arm Selection. We appeal to the Upper Confidence Bound (UCB) principle [1, 20] to select an arm from the candidate arm pool (line 18 of Algorithm 1). With the above bandit expert selection strategy, for arm a at time t, we might collect a set of admissible bandit experts $\mathcal{M}^a_t$, and each admissible bandit expert m gives us an upper confidence bound of reward estimation for arm a: $\text{UCB}_{a,t}(m) = \mathbf{x}_a^{\top}\hat{\boldsymbol{\theta}}_t(m) + B^{\theta}_{a,t}(m)$. Therefore, we need to integrate $\text{UCB}_{a,t}(m)$ from multiple bandit experts (line 16 of Algorithm 1). We propose two strategies for this purpose, each of which has its own advantages; but both of them lead to the same upper regret bound (proved in Theorem 2 of Section 3.3).

Option 1: Average ensemble. For each arm, as in principle every admissible bandit expert is 'legitimate' to give a reward prediction for this arm, we compute an average upper confidence bound of reward based on all admissible expert models: $\text{UCB}_{a,t} = \frac{1}{|\mathcal{M}^a_t|}\sum_{m\in\mathcal{M}^a_t}\text{UCB}_{a,t}(m)$. This ensemble helps reduce variance in reward estimation among the admissible bandit experts, but might be disturbed by some least qualified bandit experts.

Option 2: Lower confidence bound of badness. Although every admissible bandit expert is guaranteed to have small enough badness to make an accurate reward prediction on the chosen arm, the uncertainty of their auditors' badness estimation may differ, which introduces another level of trade-off between exploitation and exploration in bandit expert selection. By applying the Lower Confidence Bound (LCB) principle (as we want to minimize the badness of chosen bandit experts), we select a bandit expert from $\mathcal{M}^a_t$ by the LCB of its auditor's estimated badness: $\tilde{m}_{a,t} = \arg\min_{m\in\mathcal{M}^a_t}\left(|\hat{e}_{a,t}(m)| - B^{\beta}_{a,t}(m)\right)$, and compute the UCB of arm a using the selected bandit expert $\tilde{m}_{a,t}$.

Model Update. Once the feedback $(\mathbf{x}_{a_t}, r_{a_t})$ is obtained from the environment on the selected arm a_t, we update the bandit experts and auditors in this arm's admissible model set (line 20-31 in Algorithm 1). In particular, we compare the acquired feedback against each bandit expert's estimation (line 24) to decide whether we should update the bandit expert to improve its reward estimation or the bandit auditor to improve its badness estimation. This decision comes from two factors that cause the observed reward estimation error: 1) large noise from the environment; 2) the arm is actually not change-invariant. Large noise may happen, but with a very small probability, as it follows a Gaussian distribution. Define $\epsilon = \sqrt{2}\sigma\text{erf}^{-1}(\delta_1 - 1)$, in which σ is the standard deviation of the Gaussian noise in reward feedback, and $\text{erf}(\cdot)$ is the Gauss error function. Violating the condition $|\hat{r}_{a,t}(m) - r_{a,t}| \le B^{\theta}_{a,t}(m) + \Delta_L + \epsilon$ suggests that the chosen arm might not be change-invariant, such that the bandit expert should not be updated but the bandit auditor should be. This selective update strategy helps reduce erroneous observations in both bandit experts and auditors.

3.3 Regret Analysis

The accumulated (pseudo) regret of a bandit algorithm up to time T is formally defined as $R(T) = \sum_{t=1}^{T}(\mathbb{E}[r_{a^*_t,t}] - \mathbb{E}[r_{a_t,t}])$, in which a^*_t is the best arm according to the oracle at time t, and a_t is the arm selected by the algorithm to be evaluated. We first provide high probability bounds of bandit experts' reward estimation and bandit auditor's badness estimation in the following theorem.

Theorem 1 (Confidence Bounds). *We define S_{min} as the length of the shortest stationary period up to time T and $t_c(m)$ as the first change point in the environment after bandit expert m is created. When Assumption 1 is satisfied with $\Delta_H > 4\sqrt{\lambda} + \Delta_L$ and $\Delta_L \le \frac{\sigma^2\sqrt{d\lambda\ln\frac{\lambda+T}{\lambda\delta_1}}}{T}$, and $1 < \tau \le S_{min}\frac{\sigma^2\sqrt{d\lambda\ln\frac{\lambda+T}{\lambda\delta_1}}}{2\rho T}$, at time t for any $\delta_1 \in (0,1)$, $\delta_2 \in (0,1)$ and $\delta_3 = 1-(1-\delta_1)\left[(1-\delta_2)(1-\delta_1)\rho\right]^{\max\{t-t_c(m),0\}}$, with a probability at least $1-\delta_3$, for any arm a all the bandit experts in Algorithm 1 satisfy,*

$$|\mathbf{x}_a^{\top}\hat{\boldsymbol{\theta}}_t(m) - \mathbf{x}_a^{\top}\boldsymbol{\theta}^*_{t_m}| \le B^{\theta}_{a,t}(m,a) \tag{3}$$

If there is no environment change in the past τ iterations, with a probability at least $1-\delta_2$, all bandit auditors at time t satisfy,

$$|\mathbf{x}_a^{\top}\hat{\boldsymbol{\beta}}_t(m) - \mathbf{x}_a^{\top}\boldsymbol{\beta}^*_t| \le B^{\beta}_{a,t}(m,a) \tag{4}$$

And with a probability at least $1-\delta_3$, all the selected bandit experts in $\mathcal{M}^a_t$ for arm a satisfy,

$$|\mathbf{x}_a^{\top}\hat{\boldsymbol{\theta}}_t(m) - \mathbf{x}_a^{\top}\boldsymbol{\theta}^*_t| \le \Delta_L + B^{\theta}_{a,t}(m) \tag{5}$$

This theorem specifies the threshold of minimum reward change for change-sensitive arms, i.e., Δ_H, which decides whether a change is detectable by our algorithm. We can further relax the threshold to $2B^{\theta}_{a,t}(m) + 2B^{\beta}_{a,t}(m) + \Delta_L$, which is shrinking over time and related to the current model uncertainties of the bandit expert and auditor. This would allow us to recognize more subtle reward changes across different stationary periods so as to improve the model estimation quality on the fly. On the other hand, Δ_L is the threshold deciding whether an arm is change-invariant with respect to two stationary periods. It is thus the resolution of our bandit experts in recognizing the ground-truth bandit parameters of their designated periods. Note, Δ_L is a parameter of the environment, rather than a parameter of our model. The parameter τ determines the number of most recent observations used for estimating the bandit auditors. Although a larger τ naturally leads to better badness estimation quality in the auditors, it cannot be arbitrarily large as it brings in out-of-date observations to the auditors. Theorem 1 imposes an upper bound of τ to guide practical use of our algorithm. Proofs this this theorem can be found in Appendix 7.1.

In the following two lemmas, we bound the probability of false negative and false positive selection of bandit experts, which proves the validity of our designed bandit expert selection strategy.

Lemma 1. *A false negative selection happens when the bandit expert m is not selected in its designated period or for the truly change-invariant arms to it in other periods. Denote the probability of a false negative selection as P_{FN}. At time t, we have $P_{FN} = \mathbb{P}\big(|\hat{e}_{a,t}(m)| > B^{\theta}_{a,t}(m) + B^{\beta}_{a,t}(m) + \Delta_L \,\big|\, |\mathbf{x}_a^\top \boldsymbol{\theta}^*_t - \mathbf{x}_a^\top \boldsymbol{\theta}^*_{t_m}| \le \Delta_L\big) \le (1-\delta_2)(1-\delta_3)$, in which δ_2 and δ_3 are defined in Theorem 1.*

Lemma 2. *A false positive selection happens when the environment has changed and the current arm is change-sensitive to a particular bandit expert m, but the bandit expert is mistakenly selected. Denote the probability of a false negative selection as P_{FP}. When Assumption 1 holds and $\Delta_H > 4\sqrt{\lambda} + \Delta_L$, we have at time $t > t_c(m)$, $P_{FP} = \mathbb{P}\big(|\hat{e}_{a,t}(m)| \le B^{\theta}_{a,t}(m) + B^{\beta}_{a,t}(m) + \Delta_L \,\big|\, |\mathbf{x}_a^\top \boldsymbol{\theta}^*_t - \mathbf{x}_a^\top \boldsymbol{\theta}^*_{t_m}| > \Delta_L\big) \le 1 - \big((1-\delta_1)(1-\delta_2)\rho\big)^{t-t_c(m)}$, in which δ_1 and δ_2 are defined in Theorem 1.*

Intuitively, a false negative selection of bandit experts happens when both the reward and badness estimations on a change-invariant arm exceed their confidence bounds; and a false positive selection happens when a change-sensitive arm undergoes substantial reward change but both the expert's reward estimation and the auditor's badness estimation still stay within their confidence bounds. Putting all these analyses together, we have the following regret bound for our proposed algorithm.

Theorem 2 (Regret bound of DenBand). *Under the same condition as stated in Theorem 1, and $(\delta_2 + \delta_3 - \delta_2\delta_3) \le \frac{1}{2S_{max}}$, using any bandit expert in $\mathcal{M}^a_t$ for UCB-based arm selection guarantees that with a probability at least $(1-\delta_3)(1-\delta_2)$ the expected accumulated regret of DenBand up to time T can be bounded as,*

$$R(T) \le 2\Gamma_T\Big(3\sigma^2\sqrt{d\ln\frac{\lambda+U_T}{\lambda\delta_1}} + \sqrt{\lambda}\Big)\sqrt{S_{max} d \ln(\lambda + \frac{S_{max}}{d})} + (\frac{1}{\rho}+2)\sigma^2\sqrt{d\lambda\ln\frac{\lambda+T}{\lambda\delta_1}}$$

where S_{max} is the length of the longest stationary period up to time T, Γ_T is the total number of environment changes till T, and $U_T = \max\{|\mathcal{I}^{\theta}_T(m)|\}_{m\in\mathcal{M}_T}$.

U_T is the maximum number of updates taken among any of the bandit experts; and it is clearly smaller than T. Hence the accumulated regret is in the order of $O(\Gamma_T\sqrt{S_{\max}\ln T\ln S_{\max}})$, which is arguably a very tight upper regret bound without any further assumption about the environment. Detailed proof and interpretation of Theorem 2 can be found in Appendix 7.2.

Theorem 2 provides a general upper regret bound of our DenBand in the order of $O(\Gamma_T\sqrt{S_{\max}\ln T\ln S_{\max}})$. To better interpret this upper regret bound under different circumstances, we consider the following two special environment cases. **Case 1:** When the changes are evenly (or almost evenly) distributed, i.e., $S_{\max} = S = \frac{T}{\Gamma_T}$, the resulting regret bound can be rewritten as $O(\sqrt{\Gamma_T T}\ln T)$, which matches the general lower regret bound in an abruptly changing environment proved in [12] without further assumptions about the environment. **Case 2:** When the distribution of changes are highly unbalanced such that there is a single very long stationary period and many very short stationary periods, by denoting the short periods with a superscript 's', the final upper regret bound can be rewritten as $O(\sqrt{S_{\max}\ln T\ln S_{\max}} + \Gamma_T\sqrt{S^s_{\max}\ln T\ln S^s_{\max}})$ where $S^s_{\max} \ll S_{\max}$. For example, when $S^s_{\max} \ll \frac{S_{\max}}{\Gamma_T^2}$, the regret can be further upper bounded by $O(\sqrt{S_{\max}\ln T\ln S_{\max}})$, which matches and is better than the upper regret bound of running a single contextual bandit over the whole period (as $S_{\max} < T$).

Our regret analysis in those two special cases supports the validity of DenBand, and we can further generalize our analysis of it under other reward assumptions. In particular, our theoretical analysis supports that any contextual bandit algorithm can be used as a bandit expert in DenBand, as long as its reward estimation error is bounded with a high probability, which corresponds to $B^{\theta}_t(m,a)$ in Eq (3). The overall regret of DenBand will only be a factor of the actual number of changes in the environment, which is arguably inevitable without further assumptions about the environment.

4 EXPERIMENTS

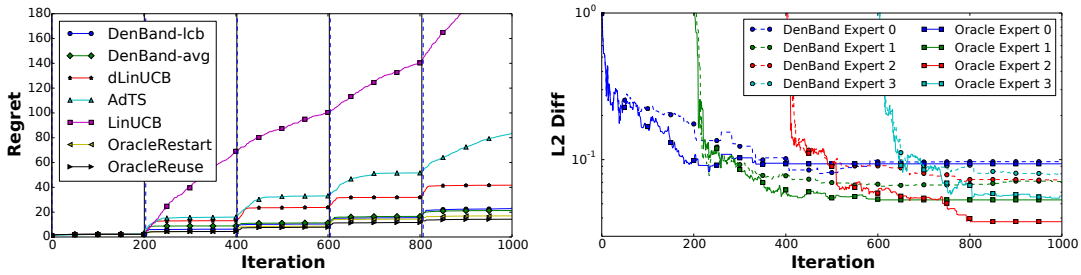
We tested DenBand on a comprehensive set of evaluation datasets, which include a synthetic dataset and three real-world recommendation datasets. On all datasets, we compared against the following bandit baselines: LinUCB [20], a reference stationary linear contextual bandit algorithm; dLinUCB [35], AdTS [14], WMDUCB1 [38], and Meta-Bandit [15], which are state-of-the-art bandit algorithms for piece-wise stationary environment. For DenBand, we tested two variants of it: DenBand-lcb, which uses lower confidence bound of badness to select a bandit expert, and DenBand-avg, which takes the average reward UCBs from all admissible bandit experts.

4.1 Experiments on Synthetic Dataset

• **Simulation Settings.** In simulation, we generate a size-K (K = 1000) arm pool $\mathcal{A}$, in which each arm a is associated with a d-dimensional (d = 10) feature vector $\mathbf{x}_a \in \mathbb{R}^d$ with $\|\mathbf{x}_a\|_2 \le 1$. Similarly, we create a set of ground-truth bandit parameters $\boldsymbol{\theta}^* \in \mathbb{R}^d$ with $\|\boldsymbol{\theta}^*\|_2 \le 1$, which are not disclosed to the learners. The standard deviation of Gaussian noise σ on the reward is set to 0.05 by default. To simulate an abruptly changing environment, after every S rounds, we randomize $\boldsymbol{\theta}^*$ with respect to the constraint that at least ρ portion of arms in $\mathcal{A}$ satisfy $|\mathbf{x}_a^\top\boldsymbol{\theta}^*_{t_{c_j}} - \mathbf{x}_a^\top\boldsymbol{\theta}^*_{t_{c_{j+1}}}| > \Delta_H$. And by default, we set S to 200, Δ_H to 0.7, and ρ to 0.2.

• **Empirical Regret Comparisons.** Under this simulation setting, we execute all algorithms 1000 iterations and report their accumulated regret in Figure 3 (a). Because LinUCB imposes a stationary assumption about the environment, it suffers almost linearly increasing regret after the first change point. Both AdTS and dLinUCB can react to the environment changes, but they are slow in doing so and thus accumulate increasing regret. Both of our proposed algorithms, DenBand-lcb and DenBand-avg, can quickly identify the changes and create corresponding bandit experts to capture the new reward distributions. The solid and dashed vertical lines in Figure 3 (a) show the actual and detected change points by DenBand-lcb respectively. We can clearly notice that DenBand can almost immediately respond to the changes in the environment. To improve visibility of Figure 3 (a), WMDUCB and Meta-Bandit are excluded.

In addition, we also include two oracle algorithms in Figure 3 (a). One is to start a new bandit expert at each change point and only uses it in this period, named as *OracleRestart*. The other, named as *OracleReuse*, maintains one bandit expert for each stationary



(a) Accumulated regret over time

(b) Parameter estimation quality of bandit experts

Figure 3: Performance comparison on a synthetic dataset.

Table 1: Accumulated regret with different settings of the non-stationary environment.

$(\sigma, \Delta_H, S, \rho)$	(0.05, 0.7, 200, 0.8)	(0.1, 0.7, 200, 0.8)	(0.05, 0.7, 400, 0.8)	(0.05, 0.5, 200, 0.8)	(0.05, 0.5, 200, 0.5)	(0.05, 0.5, 200, 0.2)
DenBand-lcb	**22.48 ± 3.12**	36.68 ± 6.64	16.55 ± 1.48	**23.24 ± 3.01**	**18.96 ± 1.96**	22.45 ± 2.32
DenBand-avg	25.54 ± 3.88	**33.87 ± 4.13**	**13.08 ± 2.12**	23.51 ± 2.55	20.40 ± 2.06	**19.87 ± 1.85**
dLinUCB	35.12 ± 2.20	54.58± 3.56	22.33 ± 2.83	35.72 ± 2.05	34.45 ± 2.83	36.65 ± 2.10
AdTS	65.16± 20.30	67.43 ± 21.49	38.61 ± 5.08	70.74 ± 26.60	58.59 ± 14.13	52.45 ± 14.21
LinUCB	253.22 ± 7.12	263.14 ± 11.76	257.50 ± 8.07	216.60 ± 11.32	206.29 ± 15.71	164.68 ± 8.98
WMDUCB1	519.73 ± 21.53	528.48 ± 23.63	473.13 ± 43.29	484.23 ± 17.73	404.17 ± 10.95	372.30 ± 9.68
Meta-Bandit	585.70 ± 3.72	585.97 ± 4.62	499.83 ± 5.14	454.56 ± 4.29	412.30 ± 2.68	362.96 ± 2.17

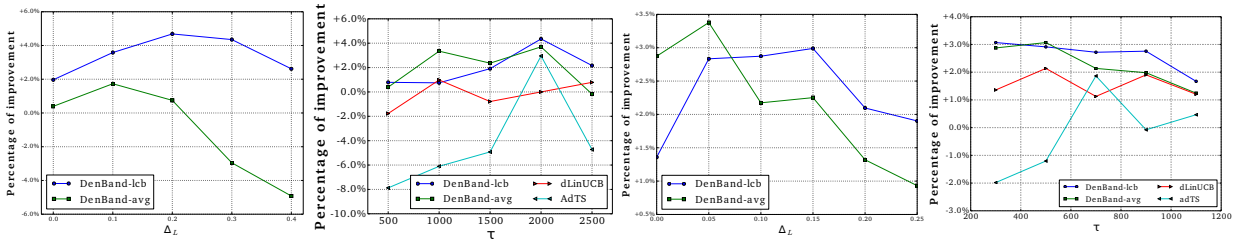
Table 2: Accumulated regret with different hyperparameter configurations.

(τ, Δ_L)	(30, 0.0)	(50, 0.0)	(80, 0.0)	(100, 0.0)	(100, 0.025)	(100, 0.05)	(100, 0.1)	(100, 0.15)
DenBand-lcb	24.36 ± 4.23	22.48 ± 3.12	19.33 ± 2.65	18.36 ± 3.27	19.27 ± 1.74	19.58 ± 1.38	21.08 ± 3.17	23.75 ± 5.96
DenBand-avg	25.73 ± 1.53	21.19 ± 3.29	18.87± 2.59	19.04 ± 1.65	19.52 ± 2.10	22.78 ± 2.19	26.84 ± 4.03	31.81 ± 6.08
dLinUCB	35.08 ± 2.59	35.12 ± 2.20	40.28 ± 2.52	42.86 ± 2.23	-	-	-	-
AdTS	94.90 ± 13.08	65.16 ± 20.30	123.05 ± 14.94	91.31 ± 20.08	-	-	-	-

period, and uses all ground-truth reusable experts for reward estimation in change-invariant arms. DenBand performs very closely to such optimal algorithms, although it does not get access to the ground-truth environment changes. Figure 3 (b) shows the parameter estimation quality (i.e., the L2 difference between the estimated parameter and the ground truth parameter) for four dynamically created bandit experts in DenBand-lcb. For comparison, we also include the estimation quality of experts from the OracleReuse algorithm. The good estimation quality of each bandit expert in DenBand further verifies our conclusion in Theorem 1 about the confidence bound of admissible bandit experts' estimation quality.

• **Sensitivity to Environment Settings.** According to our regret analysis, the performance of DenBand depends on the environment settings: including standard deviation σ in the Gaussian noise, length S of stationary period, proportion ρ of change-sensitive arms, and magnitude Δ_H of reward change. We varied them in simulation to investigate their influence on the algorithms. We ran all algorithms for 10 times and reported the mean and standard derivation of obtained accumulated regret in Table 1. DenBand consistently achieved the best performance against all baselines in all environment settings. As expected, a larger noise level σ leads to worse regret in almost all algorithms. Since T is fixed in our simulation, a smaller S leads to a larger Γ_T, which linearly scales DenBand's regret. When Δ_H becomes smaller, the regret of the two variants of DenBand and AdTS are further reduced. This is because once Δ_H satisfies the requirement in Theorem 1, the change can be confidently detected; and in the meanwhile, because of a smaller Δ_H, the added regret from using a false-positive bandit expert becomes smaller. Lastly, ρ does not seriously affect the performance of DenBand, which indicates that the algorithm is robust to ρ as long as Assumption 1 is satisfied.

• **Sensitivity to Hyper-Parameters.** To verify the robustness of our proposed algorithm, we studied the effect of two important hyper-parameters in Table 2: τ, whic determines the number of most recent observations used in bandit auditors, and Δ_L, which introduces flexibility and accounts for the existence of change-invariant arms with infinitesimal reward shift. As a comparison, we also studied the effect of τ on dLinUCB and AdTS, which also use a sliding window for change detection. From Table 2 we can find that both variants of DenBand are robust to its parameter τ as long as it falls into the required range. This result verified our theoretical analysis about the role of τ in Theorem 1, which defines an upper and lower bound of τ for the proved confidence bound. On the contrary, the baselines, especially AdTS, are very sensitive to the setting of τ. For the parameter Δ_L, DenBand-lcb is very robust to it, but DenBand-avg is not. This can be explained from two perspectives: 1). As Δ_L accounts for the existence of change-invariant arms with infinitesimal reward shift, it should be upper bounded as required in Theorem 1. 2). When Δ_L is set too large, inadmissible bandit experts may be mistakenly selected. In this case, DenBand-avg suffers from a contaminated average ensemble of bandit experts in reward prediction. DenBand-lcb reduces the risk by selecting a bandit expert by its lower confidence bound of estimated badness.



(a) Effect of Δ_L on Yahoo dataset

(b) Effect of τ on Yahoo dataset

(c) Effect of Δ_L on Snapchat dataset

(d) Effect of τ on Snapchat dataset

Figure 4: Effect of hyper-parameters on DenBand on two real-world datasets

4.2 Experiments on Real-World Datasets

• **Datasets.** We tested all algorithms on the following three real-world recommendation datasets.

The first dataset is a large collection of Yahoo frontpage recommendation log, made available by the Yahoo Webscope program [20]. It contains 45,811,883 user visits to Yahoo Today Module in a ten-day period in May 2009. Unbiased offline evaluation is performed on this dataset following the offline evaluation protocol used in [20, 21].

The second dataset is a user click log collected from the Snapchat lens recommendation platform over a one-week period. This dataset contains hundred millions of anonymous observations related to 495 unique lenses and more than five hundred thousand randomly selected users. Each observation contains an anonymous user id, a lens id which is recommended by the logging policy, and corresponding user actions on this lens. Each lens is associated with an 8-dimensional feature vector constructed by historical click statistics. The sharing-related actions on items are considered as positive user click feedback. Due to the sparsity of observations, users are grouped into 22 groups according to their demographic information, and estimate one bandit model in each user group. The same offline evaluation protocol is used here as in the Yahoo news recommendation dataset.

The third data is extracted from the music streaming service Last.fm, which is made available on the HetRec 2011 workshop [2]. This dataset contains 1892 users and 17632 items (artists). We treat the 'listened artists' in each user as positive feedback. Following the settings in [8, 32, 36], we pre-processed the dataset in order to fit them into a contextual bandit setting. And we followed [15, 35] to simulate a non-stationary environment: we ordered observations chronologically inside each user, and built a single hybrid user by merging different users. Hence, the boundary between two consecutive batches of observations from two original users is treated as the preference change of the hybrid user.

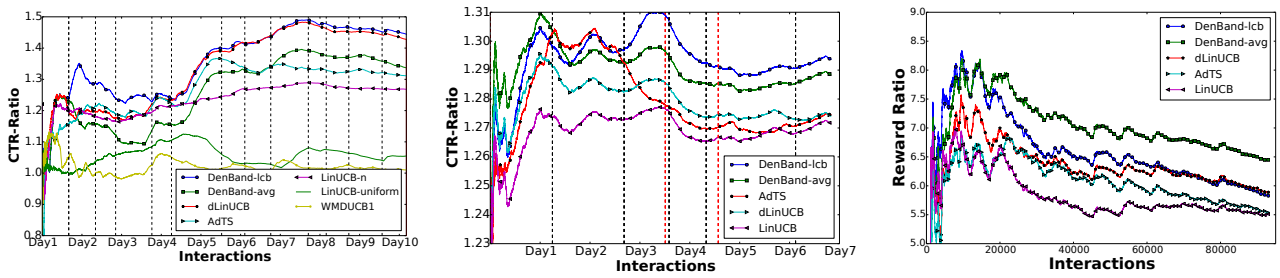
• **Sensitivity of Hyper-Parameters.** To test the proposed algorithms' sensitivity to hyper-parameters and identify the best hyper-parameter setting, we perform parameter tuning using one day's data from yahoo dataset and two days' data from Snapchat dataset. We vary Δ_L and τ, and report the relative performance improvement compared with LinUCB in Figure 4. From Figure 4 (a) and (c), we notice that DenBand-lcb is very robust to the setting of Δ_L, while the performance of DenBand-avg can be bad when Δ_L is too large. As discussed in our study of hyper-parameters in Section 4.1, the reason is that for a larger Δ_L, DenBand is more likely to include false positive models as admissible bandit experts. In this case, the reward estimation quality may become inaccurate in DenBand-avg, which evenly trust all admissible experts' output. As for τ, in Figure 4 (b) and (d) we observed that AdTS is very sensitive to the choice of τ, while both variants of DenBand are quite robust to it as long as it is within a reasonable range.

• **Recommendation Quality** In Figure 5 (a), we report normalized Click-Through Rate (CTR) from different algorithms based on the corresponding logged random strategy's CTR on the Yahoo dataset. We set τ to 2,000 and Δ_L to 0.3 based on the hyper-parameter tuning results. From Figure 5 (a), we can find that DenBand-lcb achieves significant improvement compared with all baselines, especially at the beginning of the testing period. WMDUCB1 performs the worst as it cannot utilize any available context information. We also looked into the detected change points by DenBand-lcb, and found that it detected 25 changes in total, and we plotted 12 of them in Figure 5 (a) using vertical lines (to increase visibility of the figure). We can find close correspondence between its detected change points and its performance improvement.

Using the same evaluation protocol, we report the CTR ratio between different algorithms and the logging policy on Snap dataset in Figure 5 (b). We set τ to 500 and Δ_L to 0.1. On this dataset, the context-free algorithms perform significantly worse than the contextual ones, so that we excluded them from the comparisons. The results show that our algorithms achieve a 29% - 31% improvement comparing to the logging policy, and a 3.1% improvement against LinUCB in a per-user basis. Comparing to AdTS, although our proposed algorithm performs similarly at the beginning, it caught up very quickly later. We plot the detected change points for two largest groups of users (users in the same group share the same set of bandit experts) in Figure 5 (b) using vertical lines of different colors, which well correlate with performance improvement during the adaptive recommendation process.

The ratio between reward from the bandit algorithms and that from a random selection policy on LastFM dataset is reported in Figure 5 (c), where Δ_L is set to 0.1 and τ is set to 300. From this result, we can see that DenBand, especially DenBand-avg outperform all the baselines. LinUCB performs the worst as it failed to capture the non-stationarity of the environment. Since the distribution of items is highly skewed [8], the context-free bandits perform very poorly on this dataset. We therefore decide to exclude them from comparison in the figure.

[2] http://grouplens.org/datasets/hetrec-2011



(a)Yahoo dataset (b) Snapchat dataset (c) LastFM dataset

Figure 5: Normalized CTR comparison on three real-world datasets.

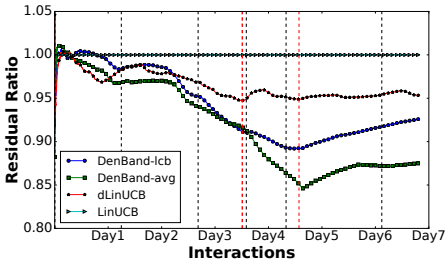


Figure 6: Residual ratio on Snapchat dataset

• Reward estimation quality and qualitative study. In order to investigate the effectiveness of DenBand in reward estimation and illustrate the source of its performance improvement, we looked into its reward prediction error on the selected items on Snapchat dataset. In Figure 6, we normalize the accumulated residuals (i.e., reward prediction error) from different algorithms by that from LinUCB. It shows that both variants of DenBand have smaller reward estimation error on the selected items. In addition, we illustrate the detected change points for the two largest user groups using vertical lines. The detected changes correlate well with the turning points of residual ratio in DenBand. For example, at the first and last detected change points, the residual ratio is increasing, which indicates that the old bandit model is becoming bad; and thus new bandit experts are created to account for the new environment.

To reveal how DenBand recognizes the changes of users' interests in a context-dependent manner, we looked into the high reward items, change-invariant items, and change-sensitive items according to our sequentially created bandit experts and their corresponding auditors on the LastFM dataset. For each bandit expert and auditor pair, we used the learnt models in the bandit expert to get the top 500 high reward items, and used its auditor to get the top 500 change-invariant items and top 500 change-sensitive items separately. As each item is associated with some short text descriptions provided by the users, we then generated word clouds for each group of those selected items to summarize the learnt bandit models in Figure 7. It is interesting to find that the change-invariant items tend to be related with geographical regions. For example, the change-invariant items in group 1 are mostly about Brazilian music, and those in group 2 are related to Japanese music. The change-sensitive items are mostly related to those more common genres of music. And the high reward items are a mix of these two types. Hence recognizing the changing and stable users' interest is essential in making satisfactory recommendations.

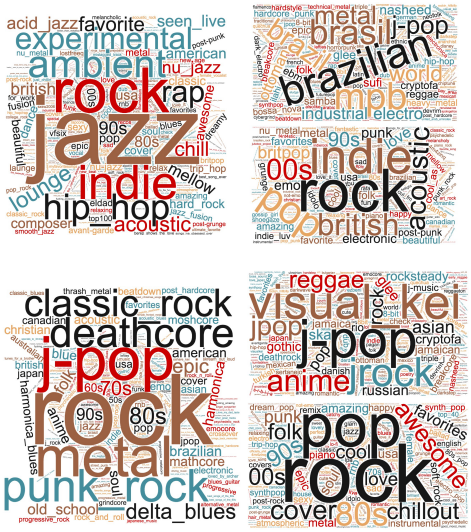


Figure 7: Word cloud visualization of two bandit experts and their auditors on LastFM. In each of the word cloud groups, the left one shows the tags from high reward items selected by the bandit expert; the upper right one shows tags from change-invariant items, and the lower right one shows tags collected from change-sensitive items according to the bandit auditor, respectively.

5 CONCLUSIONS & FUTURE WORK

We studied contextual bandits for adaptive recommendation in a piece-wise stationary environment. Capitalizing on the context-dependent property of environment changes, bandit models are dynamically ensembled and reused to conquer the non-stationary environment. Rigorous regret analysis validates the convergence of the proposed solution, and extensive empirical evaluations on simulation and three large real-world datasets verified the effectiveness and reliability of the proposed algorithm.

In this work, we treat bandit learners as independent from each other. As existing works have shed light on collaborative bandit learning [13], it is meaningful to study non-stationary bandits in a collaborative manner, e.g., monitoring the change and reusing models across users. It is also important to study a continuously changing environment, such as Brownian motion or periodical changes, where the context-dependent changes help us better maintain and create bandit models.

6 ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation Grant IIS-1553568 and IIS-1618948, and Bloomberg Data Science PhD Fellowship.

7 APPENDIX

7.1 Proof of Theorem 1

Proof of Eq (3)**:** For the bandit expert m created at time t_m, we split its training instances in $\mathcal{I}_t^\theta(m)$ up to time t into two sets $\mathcal{H}_t(m)$ and $\tilde{\mathcal{H}}_t(m)$. The instances in $\mathcal{H}_t(m)$ are all from the reward distribution governed by $\boldsymbol{\theta}^*_{t_m}$, while the instances in $\tilde{\mathcal{H}}_t(m)$ are not. According to the first order optimum condition in ridge regression and the definition of $\mathbf{A}_t(m)$, the following equation can be obtained,

$$\hat{\boldsymbol{\theta}}_t(m) - \boldsymbol{\theta}^*_{t_m} = \mathbf{A}_t^{-1}\Big(\sum_{i\in\mathcal{I}_t^\theta(m)} \mathbf{x}_i\eta_i - \lambda\boldsymbol{\theta}^*_{t_m} - \sum_{i\in\tilde{\mathcal{H}}_t(m)} \mathbf{x}_i\mathbf{x}_i^\top(\boldsymbol{\theta}^*_{t_m} - \boldsymbol{\theta}^*_i)\Big)$$

Based on the self-normalized martingale bound in Theorem 1 of [1], for any $\delta_1 \in (0,1)$, with a probability at least $1-\delta_1$, we have

$$\|\hat{\boldsymbol{\theta}}_t(m) - \boldsymbol{\theta}^*_{t_m}\|_{\mathbf{A}_t} \le \sigma^2\sqrt{d\ln\frac{\lambda + |\mathcal{I}_t^\theta(m)|}{\lambda\delta_1}} + \sqrt{\lambda} + \|\sum_{i\in\tilde{\mathcal{H}}_t(m)} \mathbf{x}_i\mathbf{x}_i^\top(\boldsymbol{\theta}^*_t - \boldsymbol{\theta}^*_i)\|_{\mathbf{A}_t^{-1}}$$

To simplify notations, we define $F_t(m) = \sum_{i\in\tilde{\mathcal{H}}_t(m)} \mathbf{x}_i\mathbf{x}_i^\top(\boldsymbol{\theta}^*_t - \boldsymbol{\theta}^*_i)$. Then we have,

$$\begin{aligned}|\mathbf{x}_a^\top\hat{\boldsymbol{\theta}}_t(m) - \mathbf{x}_a^\top\boldsymbol{\theta}^*_{t_m}| &\le \|\hat{\boldsymbol{\theta}}_t(m) - \boldsymbol{\theta}^*_{t_m}\|_{\mathbf{A}_t}\|\mathbf{x}_{a_t}\|_{\mathbf{A}_t^{-1}} \qquad (6)\\ &\le\Big(\sigma^2\sqrt{d\ln\frac{\lambda + |\mathcal{I}_t^\theta(m)|}{\lambda\delta_1}} + \sqrt{\lambda} + \|F_t(m)\|_{\mathbf{A}_t^{-1}}\Big)\|\mathbf{x}_{a_t}\|_{\mathbf{A}_t^{-1}}\end{aligned}$$

According to the definition of $F_t(m)$ and $\tilde{H}_t(m)$, $F_t(m)$ can be understood as a form of contamination in bandit expert m's estimation of the ground-truth bandit parameter $\boldsymbol{\theta}^*_{t_m}$. Denote $t_c(m)$ as the time index of the first change point after the bandit expert m is created. When $t > t_c(m)$, essentially, the contamination in $F_t(m)$ comes from two sources: First, erroneous updates from change-sensitive arms, which is referred as false positive selection of bandit experts. In Lemma 2, we proved that with a high probability there will not be any false positive selection up to time t when the auditor's estimation is not contaminated (best case scenario). But we may have additional erroneous updates when the auditor's observations contain change-sensitive arms collected after the change points. Therefore, the possible erroneous updates are at most $2\frac{(t-t_m)}{S_{\min}}\tau$. This also explains why the sliding observation window τ of the bandit auditors cannot be arbitrarily large when we explained in the requirement of this Theorem. The second source of error is the small contamination from change-invariant arms, which can be bounded by $\big(t - t_c(m)\big)\Delta_L$, due to the fact that for change-invariant arms $|\mathbf{x}_a^\top(\boldsymbol{\theta}^*_{t_m} - \boldsymbol{\theta}^*_i)| \le \Delta_L$. As a result, we have $F_t(m) \le 2\frac{t-t_m}{S_{\min}}\tau\rho + \big(t - t_c(m)\big)\Delta_L$. Thus when $\Delta_L \le \frac{\sigma^2\sqrt{d\lambda\ln\frac{\lambda+|\mathcal{I}_t^\theta(m)|}{\lambda\delta_1}}}{t-t_c(m)}$ and $\tau \le S_{\min}\frac{\sigma^2\sqrt{d\lambda\ln\frac{\lambda+|\mathcal{I}_t^\theta(m)|}{\lambda\delta_1}}}{2\rho(t-t_m)}$, we have $\|F_t(m)\|_{\mathbf{A}_t^{-1}} \le 2\sigma^2\sqrt{d\ln\frac{\lambda+|\mathcal{I}_t^\theta(m)|}{\lambda\delta_1}}$. Substituting this inequaity into Eq (6) concludes the proof.

Proof of Eq (4)**:** The training instances for $\hat{\boldsymbol{\beta}}_t(m)$ are determined by the bandit expert's estimation about the ground-truth reward distribution in the environment. Every time when $\hat{\boldsymbol{\theta}}_t(m)$ gets updated, we will revise the historical training instances in $\hat{\boldsymbol{\beta}}_t(m)$, i.e., compute $e_{a,t}(m)$ by the updated reward estimation. In addition, the bandit auditors only accumulate badness observations in the most recent τ interactions. When there is no environment change in this time window, Eq (4) can be easily derived based on [1]. For the case in which there is environment change among the τ observations, wrong selection and update of bandit experts may happen but the number is relatively small, since the number of mistakes is at most τ in each stationary period. The effect of these mistakes will be taken into account in both the bandit expert's reward estimation and the final regret in Theorem 2.

Proof of Eq (5)**:** When $t \le t_c(m)$, Eq (5) is equivalent to Eq (3). When $t > t_c(m)$, we have $|\mathbf{x}_a^\top\hat{\boldsymbol{\theta}}_t(m) - \mathbf{x}_a^\top\boldsymbol{\theta}^*_t| \le |\mathbf{x}_a^\top\hat{\boldsymbol{\theta}}_t(m) - \mathbf{x}_a^\top\boldsymbol{\theta}^*_{t_m}| + |\mathbf{x}_a^\top\boldsymbol{\theta}^*_{t_m} - \mathbf{x}_a^\top\boldsymbol{\theta}^*_t|$, in which the first term can be bounded by $B^\theta_{a,t}(m)$ according to Eq (3). With Lemma 2, when the bandit expert m is selected for arm a, with a high probability the arm is change-invariant, which means that $|\mathbf{x}_a^\top\boldsymbol{\theta}^*_{t_m} - \mathbf{x}_a^\top\boldsymbol{\theta}^*_t| \le \Delta_L$. Putting them together concludes the proof.

7.2 Proof of Theorem 2

Proof of Theorem 2. According to the reward confidence bound proved in Eq (5) of Theorem 1, and the UCB arm selection strategy in Algorithm 1, the regret at time t can be upper bounded by,

$$\begin{aligned}\mathbb{E}[r_{a^*_t,t}] - \mathbb{E}[r_{a_t,t}] &\le \mathbf{x}_{a_t}\hat{\boldsymbol{\theta}}_t(m_{a_t,t}) + B^\theta_{t,a_t}(m_{a_t,t}) + \Delta_L - \mathbf{x}_{a_t}\boldsymbol{\theta}^*_t\\ &\le 2\Delta_L + 2B^\theta_{t,a_t}(m_{a_t,t}) \qquad (7)\end{aligned}$$

Combining with the additional regret caused by the events when a bandit auditor's most recent τ observations contain environment changes, the cumulative regret can be upper bounded by $R(T) \le 2\Gamma_T\tau + 2T\Delta_L + \sum_{m\in\mathcal{M}_T}\sum_{i\in\Omega_m} 2B^\theta_{i,a_i}(m)$, in which Ω_m is the set of time indices when the bandit expert m is used for arm selection up to time T. According to the proof of Theorem 3 in [1], we have,

$$\sum_{i\in\Omega_m} B^\theta_{i,a_i}(m) \le \Big(3\sigma^2\sqrt{d\ln\frac{\lambda + |\mathcal{I}_T^\theta(m)|}{\lambda\delta_1}} + \sqrt{\lambda}\Big)\sqrt{|\Omega_m|d\ln(\lambda + \frac{|\Omega_m|}{d})}$$

Since $\sum_{i\in\Omega_m} 2B^\theta_{a_i,i}(m)$ is a concave function with respect to $|\Omega_m|$, according to the Jensen's inequality $\sum_{m\in\mathcal{M}_T}\sum_{i\in\Omega_m} 2B^\theta_{i,a_i}(m) \le |\mathcal{M}_T|\Big(3\sigma^2\sqrt{d\ln\frac{\lambda+U_T}{\lambda\delta_1}} + \sqrt{\lambda}\Big)\sqrt{S_{\max}d\ln(\lambda + \frac{S_{\max}}{d})}$, in which $U_T = \max\{|\mathcal{I}_T^\theta(m)|\}_{m\in\mathcal{M}_T}$.

Regarding to the number of bandit experts in $\mathcal{M}_T$: denote the possible number of false negative detection as k_{FN}, which follows a binomial distribution $B(T, P_{FN})$. According to Lemma 1 and the tail bound of Binomial distribution, it can be proved that with a high probability, $|\mathcal{M}_T| \le \Gamma_T + k_{FN} \le \Gamma_T + 1$. Combining all the conclusions above we have,

$$\begin{aligned}R(T) &\le 2\Gamma_T\tau + 2T\Delta_L + \sum_{m\in\mathcal{M}_T}\sum_{i\in\mathcal{S}_m} 2B^\theta_{i,a_i}(m) \le (\frac{1}{\rho} + 2)\sigma^2\sqrt{d\lambda\ln\frac{\lambda+T}{\lambda\delta_1}}\\ &+ (\Gamma_T + 1)\Big(3\sigma^2\sqrt{d\ln\frac{\lambda+U_T}{\lambda\delta_1}} + \sqrt{\lambda}\Big)\sqrt{S_T d\ln(\lambda + \frac{S_T}{d})}\end{aligned}$$

which finishes the proof. □

Proofs of Lemma 1 and Lemma 2 are omitted here due to space limit. They will be provided in a longer version of this paper.

REFERENCES

[1] Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. 2011. Improved Algorithms for Linear Stochastic Bandits. In *NIPS*. 2312–2320.

[2] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. 2011. Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web. In *Proceedings of the 3rd International Web Science Conference*. ACM, 2.

[3] Peter Auer. 2002. Using Confidence Bounds for Exploitation-Exploration Tradeoffs. *Journal of Machine Learning Research* 3 (2002), 397–422.

[4] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Mach. Learn.* 47, 2-3 (May 2002), 235–256.

[5] Dimitris Bertsimas and José Niño Mora. 2000. Restless Bandits, Linear Programming Relaxations, and a Primal-Dual Index Heuristic. *Oper. Res.* 48, 1 (Jan. 2000), 80–90.

[6] John S. Breese, David Heckerman, and Carl Kadie. *Empirical Analysis of Predictive Algorithms for Collaborative Filtering.* Technical Report MSR-TR-98-12. Microsoft Research. 18 pages.

[7] Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. 2018. Nearly Optimal Adaptive Procedure for Piecewise-Stationary Bandit: a Change-Point Detection Approach. *https://arxiv.org/abs/1802.03692* (2018).

[8] Nicolò Cesa-Bianchi, Claudio Gentile, and Giovanni Zappella. 2013. A Gang of Bandits. *In Pro. NIPS* (2013).

[9] Olivier Chapelle and Lihong Li. 2011. An Empirical Evaluation of Thompson Sampling. In *Advances in Neural Information Processing Systems*. 2249–2257.

[10] Joohyun Lee Fang Liu and Ness Shroff. 2018. A Change-Detection based Framework for Piecewise-stationary Multi-Armed Bandit Problem *(AAAI'18)*.

[11] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. 2010. Parametric bandits: The generalized linear case. In *NIPS*. 586–594.

[12] Aurélien Garivier and Eric Moulines. 2008. On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415* (2008).

[13] Claudio Gentile, Shuai Li, Purushottam Kar, Alexandros Karatzoglou, Giovanni Zappella, and Evans Etrue. 2017. On Context-Dependent Clustering of Bandits. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Doina Precup and Yee Whye Teh (Eds.), Vol. 70. PMLR, International Convention Centre, Sydney, Australia, 1253–1262.

[14] Negar Hariri, Bamshad Mobasher, and Robin Burke. 2015. Adapting to User Preference Changes in Interactive Recommendation. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*. 4268–4274.

[15] Cédric Hartland, Sylvain Gelly, Nicolas Baskiotis, Olivier Teytaud, and Michèle Sebag. 2006. Multi-armed Bandit, Dynamic Environments and Meta-Bandits. (Nov. 2006). https://hal.archives-ouvertes.fr/hal-00113668 working paper or preprint.

[16] Jaya Kawale, Hung H Bui, Branislav Kveton, Long Tran-Thanh, and Sanjay Chawla. 2015. Efficient Thompson Sampling for Online Matrix-Factorization Recommendation. In *Advances in Neural Information Processing Systems*. 1297–1305.

[17] Michal Kompan and Mária Bieliková. 2010. Content-Based News Recommendation. In *E-Commerce and Web Technologies*, Francesco Buccafurri and Giovanni Semeraro (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 61–72.

[18] Georgia Koutrika. 2018. Recent Advances in Recommender Systems: Matrices, Bandits, and Blenders. In *EDBT*.

[19] John Langford and Tong Zhang. 2008. The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information. In *NIPS*. 817–824.

[20] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of 19th WWW*. ACM, 661–670.

[21] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. 2011. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of 4th WSDM*. ACM, 297–306.

[22] Lei Li, Dingding Wang, Tao Li, Daniel Knox, and Balaji Padmanabhan. 2011. SCENE: A Scalable Two-stage Personalized News Recommendation System. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. ACM, New York, NY, USA, 125–134. https://doi.org/10.1145/2009916.2009937

[23] Wei Li, Xuerui Wang, Ruofei Zhang, Ying Cui, Jianchang Mao, and Rong Jin. 2010. Exploitation and exploration in a performance based contextual advertising system. In *Proceedings of 16th SIGKDD*. ACM, 27–36.

[24] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. 2010. Personalized news recommendation based on click behavior. In *Proceedings of the 15th international conference on Intelligent user interfaces*. ACM, 31–40.

[25] Christos H. Papadimitriou and John N. Tsitsiklis. 1999. The Complexity of Optimal Queuing Network Control. *Math. Oper. Res.* 24, 2 (May 1999), 293–305.

[26] Owen Phelan, Kevin McCarthy, Mike Bennett, and Barry Smyth. 2011. Terms of a Feather: Content-based News Recommendation and Discovery Using Twitter. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval (ECIR'11)*. Springer-Verlag, Berlin, Heidelberg, 448–459. http://dl.acm.org/citation.cfm?id=1996889.1996947

[27] Yi Qi, Qingyun Wu, Hongning Wang, Jie Tang, and Maosong Sun. 2018. Bandit Learning with Implicit Feedback. In *Advances in Neural Information Processing Systems*. 7287–7297.

[28] Vishnu Raj and Sheetal Kalyani. 2017. Taming non-stationary bandits: A Bayesian approach. *arXiv preprint arXiv:1707.09727* (2017).

[29] Herbert Robbins. 1952. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* 58, 5 (09 1952), 527–535. http://projecteuclid.org/euclid.bams/1183517370

[30] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of 10th WWW*. ACM, 285–295.

[31] Alex Slivkins and Eli Upfal. 2008. Adapting to a Changing Environment: the Brownian Restless Bandits. In *21st Conference on Learning Theory (COLT)*.

[32] Huazheng Wang, Qingyun Wu, and Hongning Wang. 2016. Learning hidden features for contextual bandits. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 1633–1642.

[33] Huazheng Wang, Qingyun Wu, and Hongning Wang. 2017. Factorization bandits for interactive recommendation. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[34] P. Whittle. 1988. Restless bandits: activity allocation in a changing world. *Journal of Applied Probability* 25, A (1988), 287–298. https://doi.org/10.1017/S0021900200040420

[35] Qingyun Wu, Naveen Iyer, and Hongning Wang. 2018. Learning Contextual Bandits in a Non-stationary Environment. In *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval (SIGIR '18)*. ACM, 495–504.

[36] Qingyun Wu, Huazheng Wang, Quanquan Gu, and Hongning Wang. 2016. Contextual Bandits in a Collaborative Environment. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 529–538.

[37] Qingyun Wu, Hongning Wang, Liangjie Hong, and Yue Shi. 2017. Returning is believing: Optimizing long-term user engagement in recommender systems. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 1927–1936.

[38] Jia Yuan Yu and Shie Mannor. 2009. Piecewise-stationary Bandit Problems with Side Observations. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*. 1177–1184.