Combinatorial Semi-Bandit in the Non-Stationary Environment

Wei Chen[1] **Liwei Wang**[2] **Haoyu Zhao**[3] **Kai Zheng**[4]

[1]Microsoft Research, Beijing, China
[2]Key Laboratory of Machine Perception, MOE, School of EECS, Center for Data Science, Peking University, Beijing, China
[3] Princeton University, NJ, USA
[4]Kuaishou Inc., Beijing, China

Abstract

In this paper, we investigate the non-stationary combinatorial semi-bandit problem, both in the switching case and in the dynamic case. In the general case where (a) the reward function is non-linear, (b) arms may be probabilistically triggered, and (c) only approximate offline oracle exists [Wang and Chen, 2017], our algorithm achieves $\tilde{O}(m\sqrt{NT}/\Delta_{\min})$ distribution-dependent regret in the switching case, and $\tilde{O}(V^{1/3}T^{2/3})$ distribution-independent regret in the dynamic case, where N is the number of switchings and V is the sum of the total "distribution changes", m is the total number of arms, and $\Delta_{\min}$ is a gap variable dependent on the distributions of arm outcomes. The regret bounds in both scenarios are nearly optimal, but our algorithm needs to know the parameter N or V in advance. We further show that by employing another technique, our algorithm no longer needs to know the parameters N or V but the regret bounds could become suboptimal. In a special case where the reward function is linear and we have an exact oracle, we apply a new technique to design a parameter-free algorithm that achieves nearly optimal regret both in the switching case and in the dynamic case without knowing the parameters in advance.

1 INTRODUCTION

Stochastic multi-armed bandit (MAB) [Auer et al., 2002a, Thompson, 1933] is a classical model that has been extensively studied in online learning and online decision making. The most simple version of MAB consists of m arms, where each arm corresponds to an unknown distribution. In each round, the player selects an arm, and the environment generates a reward of that arm from the corresponding distribution. The objective is to sequentially select the arms in each round and maximize the total expected reward. The MAB problem characterizes the trade-off between exploration and exploitation: On the one hand, one may play an arm that has not been played much before to explore whether it is good, and on the other hand, one may play the arm with the largest average reward so far to accumulate the reward.

Stochastic combinatorial multi-armed bandit (CMAB) is a generalization of the original stochastic MAB problem. In CMAB, the player may choose a combinatorial action over the arms $[m]$, and thus there may be an exponential number of actions. Each action triggers a set of arms, the outcomes of which are observed by the player. This is called the *semi-bandit* feedback. Moreover, some arms may be triggered probabilistically based on the outcome of other arms [Chen et al., 2016b, Wang and Chen, 2017, Kveton et al., 2015a,b]. CMAB has received much attention because of its wide applicability from the original online (repeated) combinatorial optimization to other practical problems, e.g. wireless networking, online advertising, recommendation, and influence maximization in social networks [Chen et al., 2013, 2016b, Wang and Chen, 2017, Gai et al., 2012, Combes et al., 2015, Kveton et al., 2014, 2015a,b,c].

All these studies focus on the stationary case, where the distribution of arm outcomes stays the same through time. However in practice, the environment is often changing. For example, in network routing, some routes are not available temporarily for maintenance; in influence maximization, student users may likely use social media less frequently during the final exam period; in online advertising and recommendation, people's preferences may change due to news events or fashion trend changes.

Motivated by such realistic settings, we consider the non-stationary CMAB problem in this paper. Let D_t denote the distribution of the arm outcomes (represented as a vector) at time t. We use two quantities, switchings and variation, to measure the changing of distributions $\{D_t\}_{t\leq T}$. The number of switchings is defined as $N := 1 + \sum_{t=2}^{T} \mathbb{I}\{D_t \neq$

Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021), PMLR 161:865–875.

$D_{t-1}\}$, and the variation is given as $V := \sum_{t=2}^{T} ||\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t-1}||_\infty$, where $\boldsymbol{\mu}_t$ is the mean outcome vector of the arms following distribution D_t. A related definition is the total variation $\bar{V} := \sum_{t=2}^{T} ||D_t - D_{t-1}||_{\text{TV}}$, where $||\cdot||_{\text{TV}}$ denotes the total variation of a distribution. The performance of the algorithm will be measured by the *non-stationary regret* instead of the regret in the stationary case.

This problem is first considered by Zhou et al. [2019], where the authors consider the non-stationary CMAB with approximation oracle but no probabilistically triggered arms. Zhou et al. [2019] only study the switching case, or the piecewise stationary case, where the non-stationarity is measured by N. Moreover, they add an assumption on the length of each stationary segment and thus bound the switchings N to be $O(\sqrt{T})$. Different from their model and assumptions, we consider the non-stationary CMAB in both the switching case (measured by N) and the dynamic case (measured by V or $\bar{V}$). We do not make assumptions on the number of switchings N and the length of stationary periods. Our contributions can be summarized as follow:

1. When we know the changing parameters N or V, we design algorithm CUCB-SW for the non-stationary CMAB problem. We show that CUCB-SW has nearly optimal distribution-dependent bound both in the switching case and the dynamic case, and the leading terms in the regret bounds are $\tilde{O}(m\sqrt{NT}/\Delta_{\min})$ and $\tilde{O}(m\sqrt{VT}/\Delta_{\min})$, where m is the total number of arms and $\Delta_{\min}$ is gap variable dependent on the distributions of arm outcomes (see Section 3 for the precise technical definition). We also show that CUCB-SW has nearly optimal distribution-independent bound in the dynamic case and the leading term in the bound is $\tilde{O}(V^{1/3}T^{2/3})$.

2. When parameters N or V are unknown, we design algorithm CUCB-BoB, which achieves sublinear regret in terms of T as long as $N < cT^\gamma$ or $V \le cT^\gamma$ for some constants c and $\gamma < 1$. Moreover, the distribution-dependent bounds in both cases and the distribution-independent bound in the dynamic case are nearly optimal when N and V are large.

3. In a special case when (a) the total reward of an action is linear in the means of arm distributions, (b) there is no probabilistically triggered arms, and (c) we have an exact oracle for the offline problem, we design ADA-LCMAB that does not need to know the parameters N or V in advance. Our algorithm has distribution-independent regret bounds $\tilde{O}(\min\{\sqrt{NT}, V^{1/3}T^{2/3} + \sqrt{T}\})$, which is nearly optimal in terms of N, V, T in both the switching case and the dynamic case.

1.1 RELATED WORKS

Multi-armed bandit Multi-armed bandit (MAB) problem is first introduced in Robbins [1952]. MAB problems can be classified into stochastic bandits and adversarial bandits. In the stochastic case, the reward is drawn from an unknown distribution, and in the adversarial case, the reward is determined by an adversary. Our model is a generalization of the stochastic case, as discussed below. The classical MAB algorithms include UCB [Auer et al., 2002a] and Thompson sampling [Thompson, 1933] for the stochastic case and EXP3 [Auer et al., 2002b] for the adversarial case. We refer to Bubeck and Cesa-Bianchi [2012] for a comprehensive coverage on the MAB problems.

Combinatorial semi-bandit Combinatorial semi-bandits (CSB) is a generalization of MAB, and there are also two types of CSB, i.e., in the adversarial or stochastic settings. Adversarial CSB was introduced in the context of shortest-path problems by György et al. [2007], and later studied extensively [Lattimore and Szepesvári, 2018]. There is also a large literature about stochastic CSB [Gai et al., 2012, Chen et al., 2016b, Combes et al., 2015, Kveton et al., 2015b]. Recently, Zimmert et al. [2019] propose a single algorithm that can achieve the best of both worlds. However, most of the previous works focus on linear reward functions. Chen et al. [2013, 2016b] initialize the study of nonlinear CSB. Chen et al. [2013] consider the problem with α-approximation oracle, and Chen et al. [2016b] generalize the model with probabilistically triggered arms, which includes the online influence maximization problem. Wang and Chen [2017] further improve the result and remove an exponential term in the regret bound by considering a subclass of CMAB with probabilistically triggered arms, and prove that the online influence maximization belongs to this subclass. Chen et al. [2016a] generalize the model in Chen et al. [2013] in another way, and they consider the CMAB problem with a general reward function that is dependent on the distribution of the arms, not only on their means.

Non-stationary bandits Non-stationary MAB can be viewed as a generalization of the stochastic MAB, where the reward distributions are changing over time. To obtain optimal regret bounds in terms of N or V, most of the studies need to use N or V as algorithmic parameters, which may not be easy to obtain in practice [Garivier and Moulines, 2011, Wei et al., 2016, Liu et al., 2018, Gur et al., 2014, Besbes et al., 2015]. Until very recently, an innovative study by Auer et al. [2019] solves the problem without knowing N or V in the bandit case and achieves optimal regret. Nearly at the same time, Chen et al. [2019] significantly generalizes the previous work by extending it into the non-stationary contextual bandit and also achieves optimal regret without any prior information, but this algorithm is far from practical. The works closest to ours are by Zhou et al. [2019] who also considers non-stationary combinatorial semi-bandits, and by Wang et al. [2019] who consider the piecewise-stationary cascading bandit. There are also some works considering non-stationary linear bandits [Russac et al., 2019, Kim and Tewari, 2019], which is a generalization of linear combinatorial bandits. However, the last two studies only achieve optimal bounds when the algorithm knows N or V. Although the algorithm in Zhou et al. [2019] is parameter-free, they make other assumptions on the length of the switching period. Moreover, they do not consider the probabilistically triggered arms.

2 MODEL

In this section, we introduce our model for the non-stationary combinatorial semi-bandit problem. Our model is derived from Wang and Chen [2017], which handles non-linear reward functions, approximate offline oracle, and the probabilistically triggering arms.

We have m base arms $[m] = \{1, 2, \ldots, m\}$. At time t, the environment samples random outcomes $\boldsymbol{X}^{(t)} = (X_1^{(t)}, X_2^{(t)}, \ldots, X_m^{(t)})$ for these arms from a joint distribution $D_t \in \mathbb{D}$. The sample random variable $X_i^{(t)}$ has support $[0, 1]$ for all i, t. Let $\mu_{i,t} = \mathbb{E}[X_i^{(t)}]$ and we use $\boldsymbol{\mu}_t = (\mu_{1,t}, \mu_{2,t}, \ldots, \mu_{m,t})$ to denote the mean vector at time t. The player does not know D_t for any t. In round $t \geq 1$, the player selects an action S_t from an action space $\mathbb{S}$ (could be infinite) based on the feedback from the previous rounds. When we play action S_t on the environment outcome $\boldsymbol{X}^{(t)}$, a random subset of arms $\tau_t \subseteq [m]$ are triggered, and the outcomes of $X_i^{(t)}$ for all $i \in \tau_t$ are observed as the feedback to the player. The player also obtains a nonnegative reward $R(S_t, \boldsymbol{X}^{(t)}, \tau_t)$ fully determined by $S_t, \boldsymbol{X}^{(t)}$ and τ_t. Our objective is to properly select actions S_t's at each round t based on the previous feedback and maximize the cumulative reward.

For the triggering set τ_t given the environment outcome $\boldsymbol{X}^{(t)}$ and the action S_t, we assume that τ_t is sampled from the distribution $D^{trig}(S_t, \boldsymbol{X}^{(t)})$, where $D^{trig}(S, \boldsymbol{X})$ is the probabilistic triggering function, and it is a probability distribution on the triggered subsets $2^{[m]}$ given the action S and environment outcome $\boldsymbol{X}$. Moreover, we use $p_i^{D,S}$ to denote the probability that action S triggers arm i when the environment triggering distribution is D. We define $\tilde{S}^D = \{i : p_i^{D,S} > 0\}$ to be the set of arms that can be triggered by action S under distribution D.

We assume that $\mathbb{E}[R(S_t, \boldsymbol{X}^{(t)}, \tau_t)]$ is a function of $S_t, \boldsymbol{\mu}_t$, and we use $r_S(\boldsymbol{\mu}) := \mathbb{E}_{\boldsymbol{X}}[R(S, \boldsymbol{X}, \tau)]$ to denote the expected reward of action S given the mean vector $\boldsymbol{\mu}$. This assumption is similar to that in Chen et al. [2016b], Wang and Chen [2017], and can be satisfied for example when variables $X_i^{(t)}$'s are independent Bernoulli random variables. Let $\text{opt}_{\boldsymbol{\mu}_t} := \sup_{S \in \mathbb{S}} r_S(\boldsymbol{\mu}_t)$ denote the maximum reward in round t given the mean vector $\boldsymbol{\mu}_t$.

The previous model is similar to that in Wang and Chen [2017], except that in this paper, we consider the non-stationary setting where D_t can change in different rounds. We assume that $\{D_t\}$ are generated *obliviously*, i.e. the generation of D_t is completed before the algorithm starts, or equivalently, the generation of D_t is independent to the randomness of our algorithm and the randomness of the previous samples $X^{(s)}, s < t$. Next, we introduce the measurement of the non-stationarity. In general, there are two measurements of the change of the environment: the first is the number of the swichings N, and the second is the variation V or $\bar{V}$. For any interval $I = [s, s']$, we define the number of switchings on I to be $N_I := 1 + \sum_{t=s+1}^{s'} \mathbb{I}\{D_t \neq D_{t-1}\}$, which can be interpreted as the number of stationary segments. As for the variation, we define $V_I := \sum_{t=s+1}^{s'} ||\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t-1}||_\infty$, which denotes the total change of the mean. By the above definitions, we have a simple fact that $V_I \leq N_I$. Another similar quantity is the total variation, and the formal definition is given as $\bar{V}_I := \sum_{t=s+1}^{s'} ||D_t - D_{t-1}||_{\text{TV}}$, where $|| \cdot ||_{\text{TV}}$ denotes the total variation of the distribution.

V is a lower bound of $\bar{V}$ (see Lemma 9 in Luo et al. [2018]). In some cases, $\bar{V}$ can be in order $\Theta(T)$ while V is a constant (just consider distribution varies but with the same expectation). In non-stationary multi-armed bandits, V is more frequently used compared with $\bar{V}$ [Gur et al., 2014, Auer et al., 2019]. $\bar{V}$ is often used in contextual bandits [Luo et al., 2018, Chen et al., 2019].

For convenience, we use N, V and $\bar{V}$ to denote $N_{[1,T]}$, $V_{[1,T]}$ and $\bar{V}_{[1,T]}$ respectively. When we use N to measure the non-stationarity, we say that we are considering the switching case. Otherwise, when we are using parameters V or $\bar{V}$, we say that we are in the dynamic case. We also define $K = \max_{t,S} |\tilde{S}^{D_t}|$ to be the maximum number of arms that can be triggered by an action in any round. Clearly, we have $K \leq m$.

Now we can introduce the measurement of the algorithm. Given an online algorithm $\mathcal{A}$, we assume that $\mathcal{A}$ has access to an offline (α, β)-approximation oracle O, which takes the input $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$ and returns an action S^{O} such that $\Pr\{r_\mu(S^{\mathsf{O}}) \geq \alpha \cdot \text{opt}_{\boldsymbol{\mu}}\} \geq \beta$. Here, α can be interpreted as the approximation ratio and β is the success probability. Based on the (α, β)-approximation oracle O, we have the following definition of (α, β)-approximation non-stationary regret:

Definition 1 ((α, β)-approximation Non-stationary Regret)**.** *The (α, β)-approximation non-stationary regret for algorithm $\mathcal{A}$ during the total time horizon T is defined as the following:*

$$Reg_{\alpha,\beta}^{\mathcal{A}} := \alpha \cdot \beta \cdot \sum_{t=1}^{T} opt_{\boldsymbol{\mu}_t} - \mathbb{E}\left[\sum_{t=1}^{T} r_{S_t^{\mathcal{A}}}(\boldsymbol{\mu}_t)\right],$$

where $S_t^{\mathcal{A}}$ is the action selected by algorithm $\mathcal{A}$ in round t.

Intuitively, the first term $\alpha \cdot \beta \cdot \sum_{t=1}^{T} \text{opt}_{\boldsymbol{\mu}_t}$ is the best we can guarantee with the total knowledge of the distributions D_t for every round t, and the second term is the expected reward selected by our algorithm $\mathcal{A}$.

Our regret bounds are in the form $\tilde{O}(N^{\gamma_1}T^{\gamma_2})$ for the switching measurement and $\tilde{O}(V^{\gamma_3}T^{\gamma_4})$ for the variation measurement. Note that if we allow the distributions D_t to change arbitrarily in every round, we cannot learn the distribution at all and there is no hope to get the non-stationary regret bound "sub-linear" in terms of T. This implies that we cannot get regret bounds with $\gamma_1+\gamma_2 < 1$ or $\gamma_3+\gamma_4 < 1$, because N and V are bounded by T and the above inequalities would lead to sublinear regrets even for arbitrary changes of D_t. Thus, the best one can hope for is to achieve regret bounds with $\gamma_1 + \gamma_2 = 1$ or $\gamma_3 + \gamma_4 = 1$. Indeed, all of our algorithms in the paper achieve such regret bounds. In this case, as long as N or V is sublinear in T, we would achieve a sublinear regret in T. Moreover, in this case, we also prefer bounds with γ_2 or γ_4 as small as possible, because it would lead to better regret bound in T as long as N or V is sublinear in T. In many cases, our algorithms do achieve the minimum possible γ_2 or γ_4, as we discuss later for each algorithm.

We make the following assumptions on the problem instance similar to those in Wang and Chen [2017], which shows that many important CMAB application instances such as influence maximization and combinatorial cascading bandit satisfy these assumptions.

Assumption 1 (Monotonicity). *For any $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$ with $\boldsymbol{\mu} \leq \boldsymbol{\mu}'$ (dimension-wise), for any action S, $r_S(\boldsymbol{\mu}) \leq r_S(\boldsymbol{\mu}')$.*

Assumption 2 (1-Norm TPM Bounded Smoothness). *For any two distributions D, D' with expectation vectors $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$ and any action S, we have*

$$|r_S(\boldsymbol{\mu}) - r_S(\boldsymbol{\mu}')| \leq B \sum_{i \in [m]} p_i^{D,S} |\boldsymbol{\mu}_i - \boldsymbol{\mu}'_i|.$$

3 GENERAL ALGORITHM FOR NON-STATIONARY CMAB

In this section, we give an algorithm for the general CMAB model defined in Section 2. We first give the algorithm (CUCB-SW) when we know that parameters N or V that measure the non-stationarity. Then, we show how to combine the CUCB-SW with the Bandit-over-Bandit Cheung et al. [2019] to get a parameter-free algorithm (CUCB-BoB).

3.1 NEARLY OPTIMAL REGRET WHEN KNOWING N OR V

In this part, we show our algorithm for the non-stationary CMAB problem when we know the parameter N or V.

Algorithm 1 Sliding Window CUCB: CUCB-SW

```
Input: $m$, Oracle O, time horizon $T$, window size $w \leq T$ ($w$ depends on $V$ or $N$, see Theorem 1)
for $t = 1, 2, 3, \ldots$ do
    $T_{i,t} \leftarrow$ number of time arm $i$ has been triggered in time $\max\{t - w + 1, 1\}, \ldots, t - 1$.
    $\hat{\mu}_{i,t} \leftarrow$ empirical mean of arm $i$ during time $t - w, \ldots, t - 1$; (1 if not triggered).
    $\rho_{i,t} \leftarrow \sqrt{\frac{3 \ln T}{2T_{i,t}}}$ ($\infty$ if $T_{i,t} = 0$)
    $\bar{\mu}_{i,t} = \min\{\hat{\mu}_{i,t} + \rho_{i,t}, 1\}$
    $S_t \leftarrow \mathsf{O}(\bar{\mu}_{1,t}, \bar{\mu}_{2,t}, \ldots, \bar{\mu}_{m,t})$
    Play action $S_t$, observe samples from triggered set.
end for
```

We apply a standard technique and get a simple algorithm CUCB-SW. Although the algorithm is simple and straightforward, the analysis is quite complicated. Our main contribution is the analysis for CUCB-SW, especially when we have the approximation oracle and the probabilistic triggering arms. We will first introduce our algorithm CUCB-SW, and then state the regret bound and give some discussions on the regret bound and proof sketch.

When we know the parameters N or V, we can apply the sliding window technique to get the result for non-stationary CMAB. The resulting algorithm is simple and included as Algorithm 1: We use CUCB [Wang and Chen, 2017] in each round, but we only consider the samples in a sliding window with size w.

Generally speaking, in each round, we compute the empirical mean of each arm in a sliding window with size w. We also compute the corresponding UCB value for each arm. Then, we use the oracle O to solve the optimization problem with the UCB value of each arm as input.

To introduce the regret bound for CUCB-SW, we need to define the gap in the non-stationary case. Formally, we have the following definition.

Definition 2 (Gap). *For any distribution D with mean vector $\boldsymbol{\mu}$. For each action S, we define the gap $\Delta_S^D := \max\{0, \alpha \cdot \mathit{opt}_{\boldsymbol{\mu}} - r_S(\boldsymbol{\mu})\}$. For each arm i, we define*

$$\Delta_{\min}^{i,t} = \inf_{S \in \mathbb{S}: p_i^{D_t,S} > 0, \Delta_S^{D_t} > 0} \Delta_S^{D_t},$$

$$\Delta_{\max}^{i,t} = \sup_{S \in \mathbb{S}: p_i^{D_t,S} > 0, \Delta_S^{D_t} > 0} \Delta_S^{D_t}.$$

We define $\Delta_{\min}^{i} = +\infty$ and $\Delta_{\max}^{i} = 0$ if they are not properly defined by the above definitions. Furthermore, we define $\Delta_{\min}^{i} := \min_{t \leq T} \Delta_{\min}^{i,t}$, $\Delta_{\max}^{i} := \max_{t \leq T} \Delta_{\max}^{i,t}$ as the minimum and maximum gap for each arm.

In the above definition, the gap $\Delta_{\min}^{i,t}, \Delta_{\max}^{i,t}$ for a fixed arm i and a fixed time is similar to the definition of gap in

Wang and Chen [2017]. However, their definition is based on a single distribution D, and in our setting, we need to generalize the definition from stationary case to dynamic case where we need to take several distributions into account. Our generalization from the stationary to the dynamic case is similar to the generalization in Garivier and Moulines [2011], which takes the minimum of the gap in each round. With the above definition, we have the following regret bound.

Theorem 1 (Regret for CUCB-SW)**.** *Choosing the length of the sliding window to be* $w = \min\left\{\sqrt{\frac{T}{V}}, T\right\}$*, we have the following distribution-dependent bound,*

$$Reg_{\alpha,\beta} = \tilde{O}\left(\sum_{i\in[m]} \frac{K\sqrt{VT}}{\Delta^i_{\min}} + \sum_{i\in[m]} \frac{K}{\Delta^i_{\min}} + mK\right).$$

If we choose the length of the sliding window to be $w = \min\left\{m^{1/3}T^{2/3}K^{-1/3}V^{-2/3}, T\right\}$*, we have the following distribution-independent bound,*

$$Reg_{\alpha,\beta} = \tilde{O}\left((mV)^{1/3}(KT)^{2/3} + \sqrt{mKT} + mK\right).$$

Note that since we have $V \leq N$, we can change the parameter from V to N in both of the regret bounds. We first look at the distribution-dependent bound. Unlike the distribution-dependent bound for the stationary MAB problem, the distribution-dependent bound here has order $\tilde{O}(\sqrt{T})$. However, the $\tilde{O}(\sqrt{T})$ term is unavoidable, since the distribution-dependent bound is lower bounded by $\Omega(\sqrt{T})$ [Garivier and Moulines, 2011]. Although Garivier and Moulines [2011] only prove the lower bound in the switching case, it also applies to the dynamic case since the switching case is a special case of the dynamic case. In this way, our distribution-dependent bound is nearly optimal in both cases in terms of V, N, and T.

As for the distribution-independent bound, the leading term in the dynamic case is $(mV)^{1/3}(KT)^{2/3}$. This term is optimal in terms of V and T and we cannot further improve the exponential term. The second term $\sqrt{mKT}$ is also necessary, since this term will be the leading term when V is very small, and the non-stationary CMAB degenerates to the original stationary CMAB problem. It is well known that $\sqrt{mT}$ is the lower bound for stationary MAB problem with m arms, so the second term is also optimal. In this way, our distribution-independent bound is nearly optimal in the dynamic case. However, the bound in the switching case is not tight. Our upper bound is $N^{1/3}T^{2/3}$ but the current upper and lower bound for non-stationary MAB is $\sqrt{NT}$ [Auer et al., 2019, Chen et al., 2019]. Designing nearly optimal regret bound for the switching case is left as future work.

The readers may find that the window lengths are not the same in the theorem for distribution-dependent/independent bounds. The different lengths are crucial to get optimal bounds since we optimize the regret bounds by the window length.

The readers may also be curious about the distribution change of the triggering probability. Note that in the model part (Section 2), we do not explicitly define the distribution change of the triggering probability. However, the change of the triggering probability can change the reward a lot. The intuition is that, although we do not define the change of the triggering probability, the triggering probability is "induced" by the distribution of the outcome of each arm (e.g., the triggering of an edge in influence maximization problem is totally determined by the propagation probability of each arm). Besides, because of the TPM bounded smoothness (Assumption 2), the regret can also be bounded. In this way, we transfer the regret due to the change of the triggering probability to the regret due to the change of the arm outcome distribution, which is also the key challenge in our proof.

Now we briefly show our proof idea to handle the probabilistically triggered arms. Like the proof in Wang and Chen [2017], we first partition the action-distribution pair S^D into groups where $G_{i,j} = \{S^D \in \mathbb{S}\times\mathbb{D} | 2^{-j} < p_i^{D,S} \leq 2^{-j+1}\}$. Generally speaking, $G_{i,j}$ includes the action-distribution pairs that S triggers arm i under distribution D with probability around 2^{-j}. Then, we define another quantity $N_{i,j,t}$ for arm i that may be triggered in group $G_{i,j}$, and it will count at time s in the sliding window ends at t if $2^{-j} < p_i^{D_s,S_s} \leq 2^{-j+1}$. Intuitively, the expected number of triggers of arm i during the sliding window can be upper-bounded by $2^{-j+1}N_{i,j,t}$ and lower bounded by $2^{-j}N_{i,j,t}$. Formally, we have the following definition for $N_{i,j,t}$.

Definition 3 (Counter)**.** *Given the sliding window size w of the algorithm, in a run of the algorithm, we define the counter $N_{i,j,t}$ as the following number*

$$N_{i,j,t} := \sum_{s=\max\{t-w+1,0\}}^{t} \mathbb{I}\left\{2^{-j} < p_i^{D_s,S_s} \leq 2^{-j+1}\right\}.$$

The first step is to relate the (α,β)-approximation non-stationary regret with the quantities $N_{i,j,t}$. All the terms related to the triggering probability can be converted to $N_{i,j,t}$. Next, we bound the formula with $N_{i,j,t}$. We show that the formula is non-increasing with respect to $N_{i,j,t}$, and we find another instance N' such that $N'_{i,j,t} \leq N_{i,j,t}$. The formula with $N'_{i,j,t}$ is easier to get regret upper bound and we use that quantity to bridge between the regret and the upper bound.

3.2 PARAMETER-FREE ALGORITHM

In this section, we introduce our parameter-free algorithm for the non-stationary CMAB problem. We combined the

Algorithm 2 CUCB with Bandit over Bandit: CUCB-BoB

```
Input: Total time horizon $T$, Block size $L$, Parameters $R = R_2 - R_1$ where $R_1 \leq r_S(\mathbf{0}) \leq r_S(\mathbf{1}) \leq R_2$.
Suppose $2^k \leq L < 2^{k+1}$. Set up an EXP3.P that has $k + 1$ arms. Arm $i$ corresponds to window size $2^i$.
for $\ell = 1, 2, \ldots, \lceil \frac{T}{L} \rceil$ do
    Set up an algorithm CUCB-SW for block $\ell$, choosing the window size according to EXP3.P.
    for $t = (\ell - 1)L + 1, \ldots, \min\{\ell L, T\}$ do
        Act according to the CUCB-SW in block $\ell$.
    end for
    $R(\ell)$ is the total reward in block $\ell$.
    Pass $\frac{R(\ell)-R_1}{R}$ to EXP3.P. // Normalize to $[0, 1]$
end for
```

Bandit-over-Bandit technique [Cheung et al., 2019] with the previous sliding window CUCB algorithm (CUCB-SW), and design our parameter-free algorithm CUCB-BoB for general non-stationary CMAB problem.

Generally speaking, the Bandit-over-Bandit technique can be summarized as follow: We first divide the total time horizon T into several segments where each segment has length L (the last segment may not). Although we do not know the non-stationary parameters N or V, we can guess N or V, or other parameters used by the algorithm when we know the parameters N or V. For example, we can guess the length of the sliding window of CUCB-SW. For two different blocks, we may run the algorithm with different guessing parameters. However, random guessing cannot have a good performance guarantee, and we use a "master bandit algorithm" to control our guessing. Whenever we complete the algorithm for a block with some guessing parameter, we feed the total reward in this block to the master bandit algorithm, and the master bandit algorithm will return us the parameter used in the next block.

In our non-stationary CMAB case, we combine the Bandit-over-Bandit technique with the previous sliding window algorithm CUCB-SW. First, we assume that we have EXP3.P algorithm for the master bandit [Bubeck and Cesa-Bianchi, 2012], which is a variant of the original EXP3 algorithm. We choose EXP3.P because it is easier to derive the regret bound since the regret of EXP3.P is bounded, while the original EXP3 only has pseudo-regret bound. Furthermore, we also assume that there exists parameters $R = R_2 - R_1$ where $R_1 \leq r_S(\mathbf{0}) \leq r_S(\mathbf{1}) \leq R_2$. This assumption aims to bound the optimal value in each round. Without this assumption, the reward in each round may be too large. Our algorithm takes L as input, which denotes the length of each block, and its proper value is given in Theorem 2. We discretize the possible sliding window size in an exponential way: The possible window size are $1, 2, 4, \ldots, 2^k$ where $2^k \leq L < 2^{k+1}$. There are $O(\log_2 L)$ number of possible window sizes in total. Then in each block, we run CUCB-SW with some window size, and we control the window size by the master EXP3.P algorithm. The only thing left is that we need to feed the reward to the EXP3.P algorithm. Here we assume that the reward in each round is bounded, and we can compute the total reward in each block and normalize it into $[0, 1]$. Please see Algorithm 2 for more details.

Theorem 2. *Suppose that there exist R_1, R_2 such that $R_1 \leq r_S(\mathbf{0}) \leq r_S(\mathbf{1}) \leq R_2$ for any $S \in \mathbb{S}$ and $R = R_2 - R_1$. Choosing $L = \sqrt{mKT}/R$, we have the following distribution-independent regret bound for $Reg_{\alpha,\beta}$,*

$$\tilde{O}\left((mV)^{\frac{1}{3}}(KT)^{\frac{2}{3}} + \sqrt{R}(mK)^{\frac{1}{4}}T^{\frac{3}{4}} + R\sqrt{mKT}\right).$$

Choosing $L = K^{2/3}T^{1/3}$, we have the following distribution-dependent regret bound

$$\tilde{O}\left(K\sqrt{\sum_{i\in[m]}\frac{TV}{\Delta^i_{\min}}} + \sum_{i\in[m]}\frac{K^{\frac{1}{3}}T^{\frac{2}{3}}}{\Delta^i_{\min}} + RK^{\frac{1}{3}}T^{\frac{2}{3}}\right).$$

In this theorem, we do not need different window lengths, since the algorithm chooses for us. However, we need different block sizes. The difference aims to optimize the sublinear term in T ($T^{3/4}$ for distribution-independent and $T^{2/3}$ for distribution-dependent). We can choose $L = \sqrt{T}$ in both cases, then the sublinear term may be worse, and we may also lose some factors in terms of m, K.

Note that since $V \leq N$, we can also replace V by N in the above regret bounds. First let's focus on the distribution-independent bound. As discussed in the previous section, $(mV)^{\frac{1}{3}}(KT)^{\frac{2}{3}}$ is nearly optimal and we can not improve this term in terms of m, V, T. The last term $R\sqrt{mKT}$ is also nearly optimal. However, the term $\sqrt{R}(mK)^{\frac{1}{4}}T^{\frac{3}{4}}$ is not optimal. Nontheless, this term is sublinear and the total regret is also sublinear in T as long as $V < cT^{\gamma}$ for some $\gamma < 1$. When we change V into N, as discussed before, there is a gap between the bound $(mN)^{1/3}(KT)^{2/3}$ and the existed lower bound $\sqrt{mNT}$. Despite of this, the total regret bound is sublinear in T if $N < cT^{\gamma}$ for some $\gamma < 1$.

As for the distribution-dependent bound, the first term is nearly optimal both in the dynamic case (measured by V) and in the switching case N. The sub-optimality comes from the second term $\sum_{i\in[m]}\frac{K^{\frac{1}{3}}T^{\frac{2}{3}}}{\Delta^i_{\min}}$. Despite this, the regret bound is "sublinear" and it is nearly optimal when N or V are large. Also, note that the first term is better than the term for fixed window size because we are guessing the best window size, which can take the gaps into account. However, in the fixed window size scenario, the gaps are unknown parameters and we can only optimize through V.

Next, we briefly show the intuition of the proof. We first have the following theorem for the performance guarantee of EXP3.P algorithm [Bubeck and Cesa-Bianchi, 2012].

Proposition 1 (Regret of EXP3.P)**.** *Suppose that the reward of each arm in each round is bounded by $0 \leq r_{i,t} \leq R'$, the number of arms is K', and the total time horizon is T'. The expected regret of EXP3.P algorithm is bounded by $O(R'\sqrt{K'T'\log K'})$.*

The general idea of the proof is to decompose the (α, β)-regret of algorithm CUCB-BoB into two parts: The first part is the regret of the algorithm CUCB-SW with the best size of sliding window; the second part is the difference between the reward of CUCB-SW with best sliding window and the reward of CUCB-BoB. The bound for the first part is given in the previous section, and we want each block to be large. Otherwise, the "best" window size cannot be reached. The second part of the regret can be bounded by the EXP3.P algorithm. If we select the length of each block as L, then each reward is at order L. There are $\log_2 T$ arms in total and the time horizon for the EXP3.P algorithm is $\frac{T}{L}$. In this way, the second term is at order $\tilde{O}(L\sqrt{T/L}) = \tilde{O}(\sqrt{TL})$, and we want L to be small for the second part. Optimizing for L, we can get the bound in Theorem 2.

There are two aspects that make designing a nearly optimal parameter-free algorithm hard. The first is the combinatorial structure of the offline problem: If we want to explore a single base arm, we may afford a large regret, and if we want to eliminate a base arm, we may affect a lot of actions. The second is the approximation oracle: It is hard to detect the non-stationarity through the reward of each round since the rewards are not accurate. A very small change in the input of the oracle may lead to a huge difference in the output of the oracle. In the next section, we show that in the restricted case of linear CMAB with exact offline oracle, we do achieve near-optimal regret.

4 NEARLY OPTIMAL ALGORITHM IN SPECIAL CASE

In this section, we propose a different algorithm that achieves nearly optimal guarantee for non-stationary linear CMAB *without* any prior information. Our algorithm is based on ADA-ILTCB$^+$ of Chen et al. [2019] designed for non-stationary contextual bandits, but adapted to Linear CMAB with exact oracles (i.e. $\alpha = \beta = 1$). In ADA-ILTCB$^+$, the algorithm works on scheduled blocks with exponentially increasing length. In each block, since there is no restart in *previous blocks*, it is safe to adopt a previously learned strategy as the underlying distribution does not change. To detect non-stationarity, the algorithm randomly triggers some replay phases with different granularities and compares the performance of each policy over these intervals. If underlying distribution changes, which will cause a gap between performances over different intervals for the same policy, the algorithm will then detect it with high probability, reset all parameters and restart.

Compared with contextual bandits, which only plays over m arms, the size of action space $\mathbb{S}$ in CMAB can be exponentially large in terms of m. Though each action in CMAB can be regarded as a policy and a base arm in contextual bandits setting, a straightforward implementation of ADA-ILTCB$^+$ [Chen et al., 2019] will cause a regret depends on $|\mathbb{S}|$, which is unsatisfactory. To deal with this issue, we make full use of semi-bandit information, and adopt classic importance weight estimator for underlying unknown linear reward $\boldsymbol{\mu}_t$ [Audibert et al., 2014, Zimmert et al., 2019]. In detail, we calculate a distribution Q over the action space $\mathbb{S}$ at each round, and play a random action S drawn from Q. For the expectation $\boldsymbol{q}$ associated with distribution Q, apparently for any $i \in [m]$, $\hat{\mu}_i = \frac{X_i}{q_i}\mathbb{I}(i \in S)$ constitutes an unbiased estimation of $\boldsymbol{\mu}$ at position i, where $\boldsymbol{X}$ is a random observation with mean $\boldsymbol{\mu}$. For some notations, we use $\mathbf{1}_S$ to represent corresponding binary m-dimensional vector of a super arm S, and $\mathbb{I}_{\{\cdot\}}$ denotes the indicator function of some event. Given an interval I, denote $\hat{\boldsymbol{\mu}}_I := \sum_{t\in I}\hat{\boldsymbol{\mu}}_t/|I|$, $\widehat{\mathrm{Reg}}_I(S) := \hat{\boldsymbol{\mu}}_I^\top \mathbf{1}_{\hat{S}_I} - \hat{\boldsymbol{\mu}}_I^\top \mathbf{1}_S$ as the empirical mean and empirical regret in this interval, where $\hat{\mu}_t$ is the empirical estimation of μ_t at time t, $\hat{S}_I := \mathrm{argmax}_{S\in\mathbb{S}}\,\hat{\boldsymbol{\mu}}_I^\top \mathbf{1}_S$. $\mathrm{Conv}(\mathbb{S})$ represents the convex hull of $\mathbb{S}$ in the vector space, and define $\mathrm{Conv}(\mathbb{S})_\nu = \{\forall \boldsymbol{x} \in \mathrm{Conv}(\mathbb{S}), s.t. \forall i \in [m], x_i \geqslant \nu\}$. Given a distribution Q over $\mathrm{Conv}(\mathbb{S})_\nu$, denote its expectation as $\boldsymbol{q} := \mathbb{E}_{S\sim Q}\mathbf{1}_S$ and define $\mathrm{Var}(Q, S) := \sum_{i\in S} 1/q_i$.

Similar to contextual bandits, we show that the solution to Follow The Regularized Leader (FTRL) with log-barrier for CMAB also satisfies some nice properties as stated in the following lemma. Besides, instead of using Frank-Wolfe or other similar algorithm adopted in stationary or non-stationary contextual bandits [Agarwal et al., 2014, Chen et al., 2019], which is unavoidable as we deal with general non-linear function, FTRL for linear combinatorial semi-bandits can be solved efficiently with time complexity in polynomial order of m and T when $\mathrm{Conv}(\mathbb{S})$ can be described by a polynomial number of constraints [Zimmert et al., 2019].

Lemma 1. *For any time interval I, its empirical reward estimation $\hat{\mu}_I$, and exploration parameter $\nu > 0$, let $\boldsymbol{q}_I^\nu$ be the solution to following optimization problem (5) with constant $C = 100$:*

$$\boldsymbol{q}_I^\nu = \underset{\boldsymbol{q}\in\mathrm{Conv}(\mathbb{S})_\nu}{\mathrm{argmax}} \;\langle \boldsymbol{q}, \hat{\boldsymbol{\mu}}_I\rangle + C\nu\sum_{i=1}^{m}\log q_i \tag{5}$$

Let Q_I^ν be the distribution over $\mathbb{N}$ such that $\mathbb{E}_{S\sim Q_I^\nu}[\mathbf{1}_S] = \boldsymbol{q}_I^\nu$, then there is

$$\sum_{S\in\mathbb{S}} Q_I^\nu(S)\widehat{\mathrm{Reg}}_I(S) \leqslant Cm\nu \tag{6}$$

$$\forall S \in \mathbb{S},\ \mathrm{Var}(Q_I^\nu, S) \leqslant m + \frac{\widehat{\mathrm{Reg}}_I(S)}{C\nu} \tag{7}$$

Algorithm 3 ADA-LCMAB

```
1: Input: confidence $\delta$, time horizon $T$, action space $\mathbb{S}$
2: Definition: $\nu_j = \sqrt{\frac{C_0}{m2^jL}}$, where $C_0 = \ln\left(\frac{8T^3|\mathbb{S}|^2}{\delta}\right)$, $L = \lceil 4mC_0 \rceil, \mathcal{B}_{(i,j)} := [\iota_i, \iota_i + 2^jL - 1]$.
3: Initialize: $t = 1, i = 1$
4: $\iota_i \leftarrow t$
5: for $j = 0, 1, 2, \ldots$ do
6:    If $j = 0$, set $Q_{(i,j)}$ as an arbitrary distribution over $\mathbb{S}$; otherwise, let $(\boldsymbol{q}^{\nu_j}_{(i,j)}, Q^{\nu_j}_{(i,j)})$ be the associated solution and distribution of equation (5) with inputs $I = \mathcal{B}_{(i,j-1)}$ and $\nu = \nu_j$
7:    $\mathcal{E} \leftarrow \emptyset$
8:    while $t \leqslant \iota_i + 2^jL - 1$ do
9:       Draw REP $\sim$ Bernoulli $\left(\frac{1}{L} \times 2^{-j/2} \times \sum_{k=0}^{j-1} 2^{-k/2}\right)$
10:      if REP $= 1$ then
11:         Sample $n$ from $\{0, \ldots, j-1\}$ s.t. $\Pr[n = b] \propto 2^{-b/2}$
12:         $\mathcal{E} \leftarrow \mathcal{E} \cup \{(n, [t, t + 2^nL - 1])\}$
13:      end if
14:      Let $\mathcal{N}_t := \{n | \exists I$ such that $t \in I$ and $(n, I) \in \mathcal{E}\}$
15:      If $\mathcal{N}_t$ is empty, play $S_t \sim Q^{\nu_j}_{(i,j)}$; otherwise, sample $n \sim$ Uniform$(\mathcal{N}_t)$, and play $S_t \sim Q^{\nu_n}_{(i,n)}$
16:      Receive $\{X^t_i | i \in S_t\}$ and calculate $\hat{\boldsymbol{\mu}}_t$ according to equation (9)
17:      for $(n, [s, s']) \in \mathcal{E}$ do
18:         if $s' = t$ and ENDOFREPLAYTEST$(i, j, n, [s, t]) = Fail$ then
19:            $t \leftarrow t + 1, i \leftarrow i + 1$ and return to Line 4
20:         end if
21:      end for
22:      if $t = \iota_i + 2^jL - 1$ and ENFOFBLOCKTEST$(i, j) = Fail$ then
23:         $t \leftarrow t + 1, i \leftarrow i + 1$ and return to Line 4
24:      end if
25:   end while
26: end for
```

Procedure: ENDOFREPLAYTEST$(i, j, n, \mathcal{A})$:
Return *Fail* if there exists $S \in \mathbb{S}$ such that any of the following inequalities holds:

$$\widehat{\text{Reg}}_{\mathcal{A}}(S) - 4\widehat{\text{Reg}}_{\mathcal{B}(i,j-1)}(S) \geqslant 34mK\nu_n \log T \tag{1}$$

$$\widehat{\text{Reg}}_{\mathcal{B}(i,j-1)}(S) - 4\widehat{\text{Reg}}_{\mathcal{A}}(S) \geqslant 34mK\nu_n \log T \tag{2}$$

Procedure: ENDOFBLOCKTEST(i, j):
Return *Fail* if there exists $k \in \{0, 1, \ldots, j-1\}$ and $S \in \mathbb{N}$ such that any of the following inequalities holds:

$$\widehat{\text{Reg}}_{\mathcal{B}(i,j)}(S) - 4\widehat{\text{Reg}}_{\mathcal{B}(i,k)}(S) \geqslant 20mK\nu_k \log T \tag{3}$$

$$\widehat{\text{Reg}}_{\mathcal{B}(i,k)}(S) - 4\widehat{\text{Reg}}_{\mathcal{B}(i,j)}(S) \geqslant 20mK\nu_k \log T \tag{4}$$

With above FTRL oracle, our full implementation for non-stationary linear combinatorial semi-bandits is detailed in Algorithm 3. According to Line 15 and our estimation method, we know the expectation vector of our sampling strategy and estimated vector $\hat{\mu}_t$ are calculated as:

$$\boldsymbol{q}_t = \boldsymbol{q}^{\nu_j}_{(i,j)} \mathbb{I}_{N_t=\emptyset} + \frac{1}{|N_t|} \sum_{n \in N_t} \boldsymbol{q}^{\nu_n}_{(i,n)} \mathbb{I}_{N_t \neq \emptyset} \tag{8}$$

$$\hat{\mu}_{t,i} = \frac{X^t_i}{q_{t,i}} \mathbb{I}(i \in S_t), \quad \forall i \in [m] \tag{9}$$

For two procedures of non-stationary test in Algorithm 3, as we consider linear CMAB and have an exact oracle, which is equivalent to an Empirical Risk Minimization oracle (i.e. giving empirical loss function returns corresponding best super arm), we can use the same technique as in Chen et al. [2019] to solve two procedures with only six oracle calls.

Since a super arm is pulled at each round for CMAB, it will cause larger variance compared with pulling a single arm in contextual bandits, which requires some additional analysis. Besides, as there is no context in CMAB, we can obtain much smaller constants in ADA-LCMAB compared with original ADA-ILTCB$^+$ [Chen et al., 2019]. Now, we state the theoretical guarantee of our proposed algorithm for non-stationary linear CMAB.

Theorem 3. *Algorithm 3 guarantees Reg$^{\mathcal{A}}_{1,1}$ is upper bounded by*

$$\tilde{O}\left(\min\left\{\sqrt{mK^2NT}, \sqrt{mK^2T} + K(m\bar{V})^{\frac{1}{3}}T^{\frac{2}{3}}\right\}\right).$$

Note that in the previous theorem, the regret upper bound is nearly optimal in terms of m, N, T and $m, \bar{V}, T$. Because we know that the regret lower bound for stationary MAB problem is $\Omega(\sqrt{mT})$ with m arms, we can construct special cases to achieve regret lower bound $\Omega(\sqrt{mNT})$ in the switching case, and $\Omega((m\bar{V})^{1/3}T^{2/3})$ in the dynamic case. The technique is standard and we refer Gur et al. [2014] for more details on the construction of the special cases. However, the dependent on K may not be tight, and we left it as a future work item to tighten the dependency on K.

Another possible improvement is to change the measurement $\bar{V}$ in the regret bound into V. Although in the special cases we construct for the lower bound, V and $\bar{V}$ are at the same order, in other cases V is just a lower bound on $\bar{V}$. Improving $\bar{V}$ into V is also left as future work.

5 CONCLUSION AND FURTHER WORKS

In this paper, we study combinatorial semi-bandit (CSB) in the non-stationary environment, an extension of classic multi-armed bandits (MAB). Our CSB setting also allows non-linear reward function, probabilistically triggering behavior, and approximation oracle, which make our problem more difficult compared with non-stationary MAB or linear bandits. We first propose an optimal algorithm that achieves $\tilde{O}(m\sqrt{NT}/\Delta_{\min})$ distribution-dependent regret in the switching case and $\tilde{O}(V^{1/3}T^{2/3})$ distribution-independent regret in the dynamic case, when N or V is known. To get rid of parameter N or V, We further design a parameter-free version with regret bound $\tilde{O}(\sqrt{mNT/\Delta_{\min}} + T^{2/3}/\Delta_{\min})$ and $\tilde{O}(V^{1/3}T^{2/3} + T^{3/4})$ respectively. For a special case where the reward function is linear and we have an exact oracle, we design an optimal parameter-free algorithm that achieves nearly optimal regret both in the switching case and in the dynamic case.

As mentioned in Section 3 and 4, there are several interesting further works. The most important one is to design an optimal parameter-free algorithm for our general CSB. Second, we mainly focus on the dependence on N, V or $\bar{V}$, and T, How to improve the dependence on K is a meaningful direction. Finally, a tight lower bound in terms of all the above parameters is necessary for a full understanding of this problem.

Author Contributions

The authors are listed in alphabetic order.

Acknowledgements

This work was supported by Key-Area Research and Development Program of Guangdong Province (No. 2019B121204008)], National Key R&D Program of China (2018YFB1402600), BJNSF (L172037) and Beijing Academy of Artificial Intelligence.

References

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646, 2014.

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2014.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002a.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002b.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, pages 138–158, 2019.

Omar Besbes, Yonatan Gur, and Assaf J. Zeevi. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244, 2015. doi: 10.1287/opre.2015.1408.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012. doi: 10.1561/2200000024.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework, results, and applications. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

Wei Chen, Wei Hu, Fu Li, Jian Li, Yu Liu, and Pinyan Lu. Combinatorial multi-armed bandit with general reward functions. In *Advances in Neural Information Processing Systems*, pages 1659–1667, 2016a.

Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17(50):1–33, 2016b. A preliminary version appeared as Chen, Wang, and Yuan, "combinatorial multi-armed bandit: General framework, results and applications", ICML'2013.

Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 696–726, Phoenix, USA, 25–28 Jun 2019. PMLR.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1079–1087. PMLR, 16–18 Apr 2019.

Richard Combes, Mohammad Sadegh Talebi Mazraeh Shahi, Alexandre Proutiere, et al. Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems*, pages 2107–2115, 2015.

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

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *Algorithmic Learning Theory - 22nd International Conference, ALT 2011, Espoo, Finland, October 5-7, 2011. Proceedings*, pages 174–188, 2011. doi: 10.1007/978-3-642-24412-4_16.

Yonatan Gur, Assaf J. Zeevi, and Omar Besbes. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 199–207, 2014.

A. György, T. Linder, G. Lugosi, and G. Ottucsák. The on-line shortest path problem under partial monitoring. *The Journal of Machine Learning Research*, 8:2369–2403, 2007.

Baekjin Kim and Ambuj Tewari. Near-optimal oracle-efficient algorithms for stationary and non-stationary stochastic linear bandits. *arXiv preprint arXiv:1912.05695*, 2019.

Branislav Kveton, Zheng Wen, Azin Ashkan, Hoda Eydgahi, and Brian Eriksson. Matroid bandits: Fast combinatorial optimization with learning. *arXiv preprint arXiv:1403.5045*, 2014.

Branislav Kveton, Csaba Szepesvári, Zheng Wen, and Azin Ashkan. Cascading bandits: learning to rank in the cascade model. In *Proceedings of the 32th International Conference on Machine Learning*, 2015a.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Combinatorial cascading bandits. *Advances in Neural Information Processing Systems*, 2015b.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pages 535–543, 2015c.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint*, page 28, 2018.

Fang Liu, Joohyun Lee, and Ness B. Shroff. A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3651–3658, 2018.

Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018.*, pages 1739–1776, 2018.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58(5):527–535, 09 1952.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems*, pages 12017–12026, 2019.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Lingda Wang, Huozhi Zhou, Bingcong Li, Lav R Varshney, and Zhizhen Zhao. Be aware of non-stationarity: Nearly optimal algorithms for piecewise-stationary cascading bandits. *arXiv preprint arXiv:1909.05886*, 2019.

Qinshi Wang and Wei Chen. Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. In *Advances in Neural Information Processing Systems*, pages 1161–1171, 2017.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Tracking the best expert in non-stationary stochastic environments. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3972–3980, 2016.

Huozhi Zhou, Lingda Wang, Lav R Varshney, and Ee-Peng Lim. A near-optimal change-detection based algorithm for piecewise-stationary combinatorial semi-bandits. *arXiv preprint arXiv:1908.10402*, 2019.

Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *International Conference on Machine Learning*, pages 7683–7692, 2019.